

电子科技大学  
UNIVERSITY OF ELECTRONIC SCIENCE AND TECHNOLOGY OF CHINA

# 专业学位硕士学位论文

MASTER THESIS FOR PROFESSIONAL DEGREE



论文题目 面向多元关系的链路预测方法研究

专业学位类别	电子信息
学    号	202022080503
作者姓名	蒋育韬
指导教师	张栗粽    副教授
学    院	计算机科学与工程学院

分类号 TP391 密级 公开  
UDC <sup>注1</sup> 004.8

# 学 位 论 文

## 面向多元关系的链路预测方法研究

(题名和副题名)

蒋育韬

(作者姓名)

指导教师 张栗粽 副教授  
电子科技大学 成 都  
(姓名、职称、单位名称)

申请学位级别 硕士 专业学位类别 电子信息  
专业学位领域 计算机技术  
提交论文日期 2023 年 3 月 17 日 论文答辩日期 2023 年 5 月 15 日  
学位授予单位和日期 电子科技大学 2023 年 6 月  
答辩委员会主席 \_\_\_\_\_  
评阅人 \_\_\_\_\_

注 1: 注明《国际十进分类法 UDC》的类号。

# **Research on N-ary Relational Facts for Link Prediction**

A Master Thesis Submitted to  
University of Electronic Science and Technology of China

Discipline **Electronic Information**

Student ID **202022080503**

Author **Jiang Yutao**

Supervisor **A.Prof. Zhang Lizong**

School **School of Computer Science and Engineering**

## 摘要

相较传统的知识图谱二元关系链路预测，多元关系的链路预测是对包含多个实体与一个或多个关系构成的复杂关系事实的预测，其能够直接利用知识图谱中的复杂关系事实进行学习，减少了信息丢失与知识歧义，提高了预测结果的准确性。然而，现有方法中存在对多元关系事实理解不充分、信息利用不足的问题，降低了预测结果的准确性与方法的实用性。为此，本文从多元关系事实的信息表达以及事实信息的挖掘利用两个方向展开研究。主要的研究内容如下：

（1）提出了基于多元关系异构图表示的图变换网络链路预测模型。针对现有方法中存在忽视关系事实的异构性问题，提出了多元关系事实的异构图表示方法，并在已有关系实体对内部关联、主体三元组与辅助关系实体对之间关联的基础上增加了辅助关系实体对之间关联，增强对关系事实信息的表达。通过引入了一种图变换神经网络，实现了对关系事实本身异构交互信息的有效提取和对实体、关系的预测。通过在数据集 JF17K 上与基于一组关系实体对表示的先进模型 tNaLP<sup>+</sup>等开展对比实验，结果表明本方法在 MRR 指标上，混合关系预测提升了 0.52%，混合实体预测提升了 1.36%。

（2）提出了基于轻量化聚合的增强图变换链路预测模型。针对图变换网络因受不同长度的关系事实影响在深层网络上存在信息丢失问题，提出了基于跨通道图最大池化的轻量化聚合方法，在前面图变换网络链路预测模型基础上，将所有网络层按通道对齐后，采用图最大池化聚合各层上关键特征，实现了对输出层上关键特征信息补全。与基于主体三元组与辅助关系实体对表示的先进模型 HINGE 等相比较，实验结果显示本方法在 MRR 指标上，三元组中主关系预测提升了 1.37%，主实体预测提升了 13.88%。

（3）构建了面向公众人物信息查询的多元关系链路预测系统。本文结合社交网络场景中对于公众人物信息查询的需求，设计搭建了具备实体预测、关系预测、数据管理和历史查询功能的系统，证明了本文方法的实用性。

**关键词：**知识图谱，链路预测，多元关系，图变换网络

## ABSTRACT

Different from the link prediction on traditional binary relationships, the link prediction on N-ary relational facts or hyper-relational facts is the prediction of complex relationship facts containing multiple entities with one or more relations, which can directly learn the complex facts within the knowledge graph, reducing information loss and knowledge ambiguity, and then improve the accuracy of prediction results. However, the existing methods suffer from inadequate understanding of multivariate relational facts and insufficient information utilization, which reduces the accuracy of prediction results and the practicality of the methods. For this reason, this paper conducts research in two directions: representation of N-ary relational fact and its mining and utilization of factual information.

The main research contents are as follows.

1. A link prediction model is proposed based on a heterogeneous graph representation of N-ary relational facts with a graph transformer network. Aiming at the problem of ignoring fact heterogeneity, the thesis proposes a heterogeneous graph representation of N-ary relational facts. In addition to the intra-interaction in relation-entity pairs and the inter-interaction between primary triple and auxiliary relation-entity pairs, it adds the inter-interaction between auxiliary relation-entity pairs to enhance the representation of N-ary relational facts. Then, the thesis introduces a graph transformer neural network, which effectively extracts heterogeneous interactions within relationship facts and predicts both entities and relations. The model conducts comparison experiments on the commonly used public dataset JF17K with advanced prediction models based on a set of relation-entity pairs representation, such as tNaLP<sup>+</sup>. The results show it improves on MRR metrics with mixed-relation prediction by 0.52% and mixed-entity prediction by 1.36%.

2. An enhanced graph transformer prediction model based on lightweight aggregation is proposed. Aiming at the problem of information loss on graph transformer deep layers, which is caused by variable lengths of relational facts, the thesis proposes a lightweight aggregation method based on cross-channel graph maximum pooling. Based on the previous link prediction model of the graph transformer network, a graph maximum pooling is used to aggregate key features on each layer after aligning all

network layers by channel, which achieves the information complement on the output layer. The model conducts comparison experiments on dataset JF17K with the advanced models based on primary triple and auxiliary relation-entity pairs representation, such as HINGE. The results show it improves on MRR metrics with primary-relation prediction by 1.37% and primary-entity prediction by 13.88%.

3. An N-ary relational link prediction system for public figure information query is constructed. This paper demonstrates the practicality of this approach by designing and building a system with entity prediction, relation prediction, data management and historical query functions in conjunction with the demand for public figure information query in social network scenarios.

**Keywords:** Knowledge Graph, Link Prediction, N-ary Relation, Graph Transformer Network

# 目 录

第一章 绪论.....	1
1.1 研究工作的背景与意义.....	1
1.2 国内外研究历史与现状.....	3
1.2.1 传统二元关系的链路预测方法.....	3
1.2.2 基于多元关系的链路预测方法.....	4
1.3 主要贡献与创新.....	5
1.4 论文结构安排.....	6
1.5 本章小结.....	7
第二章 相关技术与理论基础 .....	8
2.1 引言.....	8
2.2 相关理论.....	8
2.2.1 知识图谱多元关系链路预测.....	8
2.2.2 多元关系链路预测任务中的卷积神经网络 .....	10
2.2.3 面向知识图谱的异构图神经网络.....	11
2.3 相关技术.....	12
2.3.1 图采样.....	12
2.3.2 图池化.....	13
2.3.3 面向多元关系的链路预测.....	14
2.4 主要数据集介绍.....	15
2.4.1 JF17K .....	15
2.4.2 WD50K .....	16
2.5 本章小结.....	16
第三章 基于多元关系异构图表示的图变换网络预测方法研究 .....	17
3.1 引言.....	17
3.2 基于一组关系实体对表示的相关方法分析 .....	19
3.3 基于多元关系异构图表示的图变换网络预测模型 .....	21
3.3.1 总体模型框架.....	21
3.3.2 输入子图生成过程.....	22
3.3.3 图特征提取与聚合过程.....	25
3.3.4 目标节点预测过程.....	29
3.4 实验设计与结果分析.....	29
3.4.1 实验环境与数据.....	29
3.4.2 实验评估指标.....	30
3.4.3 实验设计与参数设置.....	31

3.4.4 实验结果与分析.....	31
3.5 本章小结.....	36
第四章 基于轻量化聚合的增强图变换链路预测方法研究 .....	37
4.1 引言.....	37
4.2 基于主体三元组与辅助关系实体对及其他表示的相关方法分析 .....	39
4.3 基于轻量化聚合的增强图变换链路预测模型 .....	40
4.3.1 总体模型框架.....	40
4.3.2 轻量化聚合模块.....	41
4.4 实验设计与结果分析.....	43
4.4.1 实验环境与数据.....	43
4.4.2 实验评估指标.....	43
4.4.3 实验设计与参数设置.....	43
4.4.4 实验结果与分析.....	45
4.5 本章小结.....	49
第五章 多元关系事实预测系统设计与实现 .....	51
5.1 需求分析.....	51
5.2 总体设计.....	52
5.3 系统实现.....	54
5.3.1 系统控制模块.....	54
5.3.2 数据处理模块.....	55
5.3.3 目标预测模块.....	55
5.3.4 前端演示模块.....	56
5.4 系统展示.....	56
5.4.1 实体预测界面.....	56
5.4.2 关系预测界面.....	57
5.4.3 数据管理界面.....	57
5.4.4 历史记录界面.....	57
5.5 系统测试.....	61
5.6 本章小结.....	63
第六章 总结与展望 .....	64
6.1 全文总结.....	64
6.2 未来展望.....	65
参考文献.....	66



## 第一章 绪论

当前,知识图谱已经大量应用于学术与工业场景,并反过来促进了从各种数据源实现大规模信息抽取,以及构建和完善知识图谱与知识库的相关研究工作。链路预测是知识图谱研究中一个炙手可热的研究领域,并受益于机器学习和深度学习技术的爆发式发展而得到了迅速发展。本章首先分析了链路预测以及面向多元关系的链路预测技术的研究背景与意义,之后梳理了多元关系链路预测领域近些年来国内外的研究现状,接着针对面向多元关系链路预测技术当前所面临的困难提出了本文的创新点与解决方法,最后介绍了全文的组织结构。

### 1.1 研究工作的背景与意义

多元关系链路预测是针对知识图谱传统二元关系链路预测在处理图谱中复杂关系事实表现不佳的问题而新提出的面向复杂关系(包括二元关系)的一个研究方向。知识图谱是对真实世界信息的一种结构化表示,其本质是结构化的语义知识库,因此有时也用知识图谱来泛指各类大型知识库。知识图谱由实体和关系组成,利用实体来描述真实世界中的对象或者抽象概念,利用关系来描述实体之间的复杂关联,从而实现对现实世界中事物及其相互之间关系的形式化描述。随着互联网的蓬勃发展,网络数据信息呈现爆炸式增长,这些信息种类繁多、质量参差不齐,传统的方法已经无法满足人们迅速获取所需信息和全面掌握信息资源的需求。Google公司率先于2012年提出了知识图谱的概念,并在其搜索结果中加入了知识图谱功能,从而提升搜索引擎能力,增强了用户搜索质量和搜索体验。随后,学术与工业领域迅速对知识图谱展开了大量研究和开发工作,并推出了许多著名的知识图谱,如YAGO、Satori等,这些知识图谱包括FreeBase、WikiData、DBPedia等已成为新兴应用研究的重要数据源,并为基于知识图谱的上游应用,包括智能搜索、知识问答、智能推荐系统等提供了基础支撑。

尽管知识图谱为网络数据提供了更为合理有序的组织机制,方便用户以及应用更加快速准确地访问所需知识信息,帮助实现后续的知识挖掘与推理决策,但现实中由于知识的开放性,大型知识图谱均存在信息不完备问题,即知识的开放性与稀疏性,使得构建的知识图谱一般无法包含全部的实体与关系信息,例如FreeBase中有超过70%的人物词条中缺少出生地信息,有超过99%的缺少种族信息<sup>[2]</sup>。补全知识图谱中所缺失信息的任务叫做知识图谱补全,该任务一般可通过两种方法实现:一种是从外部信息源中抽取新的知识补充到该目标图谱中;另一种是根据该图

谱现有的知识信息进行推理,得到未包含在图谱中的新的知识进行补充,这种技术方法就叫做链路预测。通过链路预测推理出图谱中缺失的实体、关系,实现对图谱的完善,从而改善上游问答、推荐系统的实际效果,提升用户体验。因此,研究链路预测技术具有重要意义。

现实知识图谱中除了二元关系事实,同样普遍存在包含多个(超过两个)实体的多元关系事实。现有的大部分链路预测相关研究与所提方法均基于三元组结构表示的二元关系来实现,这种由头、尾实体或者实体、属性与属性值组成的三元组结构常见于各大知识图谱的关系事实之中。然而,例如多元关系事实“玛丽.居里、皮埃尔.居里和安东尼.亨利.贝克勒尔在 1903 年共同获得了诺贝尔物理学奖”所对应的包含“玛丽.居里”、“皮埃尔.居里”、“安东尼.亨利.贝克勒尔”、“1903”、“诺贝尔物理学奖”的典型五元关系事实,也同样广泛存在于各知识图谱中。Wen 等人<sup>[9]</sup>指出,FreeBase 中实际超过 1/3 的实体参与了构成其中的多元关系事实。

多元关系在对关系事实的缺失实体或链路进行预测的过程中能够提供更丰富具体的特征,也增加了处理难度,提高了对链路预测模型的要求。由于多元关系事实由两个以上实体以及不止一个关系构成,其中关系表示实体之间存在的直接关联,因此多元关系事实能够表达更加复杂的意义。传统的基于二元关系的链路预测方法无法直接对其编码预测,需要通过形式转化,将多元关系拆解成二元关系,但转化过程中容易出现两类问题,一是存在信息丢失,即舍弃多元关系事实中的部分信息从而得到一个三元组;二是存在虚拟实体或者关系,即将关系事实中的部分信息强行糅合作为一个新的实体或者关系,重新与其他实体、关系分别构成三元组,这些虚拟节点一方面不符合语义结构,另一方面也与现实知识图谱不相容,因而影响最终的预测结果。同时,多元关系链路预测存在关系事实中实体关系更多、数量不定、语义结构更加复杂、数据更加贴近现实知识图谱的问题。

现有的面向多元关系的链路预测方法根据研究者对多元关系的理解可以大致分为多元关系的一组关系实体对表示、多元关系的主体三元组与辅助关系实体对表示以及其他表示。然而不论是哪类表示,这些方法对于多元关系事实的理解仍然不够充分,或者忽略了部分关系之间的信息交互,或者忽略了关系事实的异构性特征,此外还有部分方法存在设计缺陷,只能对特定位置实体进行预测,一定程度上降低了方法的实用性。

基于上述背景,本文在倾向于多元关系的主体三元组与辅助关系实体对表示的条件下,针对前面提出的三类主要问题,从充分挖掘多元关系事实本身异构信息、方法设计既适用于实体预测也适用于关系预测的两个角度开展研究,提出了基于多元关系异构图表示的图变换网络链路预测方法以及基于轻量化聚合的增强图变

换链路预测方法,在不引入外部信息条件下,充分利用多元关系事实信息,实现了关系实体预测。对比实验结果证明了本文所提出方法的可行性和有效性。

## 1.2 国内外研究历史与现状

知识图谱链路预测问题可分为二元关系的链路预测与多元关系的链路预测。从预测的内容上,两者都包括实体预测和关系预测,即对于一个三元组(头实体、关系、尾实体),其中若是头实体或者尾实体缺失便是实体预测,而如果是关系缺失便是关系预测。到目前为止,学术界对于链路预测的研究仍主要集中在二元关系的链路预测,研究对象是一组三元组结构的关系事实。但近年逐渐有研究开始关注到知识图谱中同样普遍的多元关系,并对其研究并提出了一些链路预测方法。本节分别从这两个角度介绍链路预测的研究现状。

### 1.2.1 传统二元关系的链路预测方法

二元关系链路预测方法中包括基于张量的方法、基于翻译的方法和基于神经网络的方法<sup>[1]</sup>。基于张量的方法中,典型的是 RESCAL 模型<sup>[71]</sup>,将一个知识图谱看作是一个三维张量,模型通过最小化张量的重构误差学习到关系实体嵌入,然后将由关系实体嵌入得到的重构张量用于链路预测。类似的还有基于张量分解的 ComplEx 模型<sup>[39]</sup>,给每个关系构造了一个矩阵,该矩阵被分解并同 RESCAL 一样通过最小化重构误差进行优化。不同之处在于 ComplEx 模型采用复数值来定义关系实体嵌入,从而更有效地处理非对称关系。

基于翻译的方法主要由 TransE 模型<sup>[70]</sup>发展而来。该方法提出了一个基本假设,即一个正确的三元组表示二元关系可看作是头实体通过关系翻译得到尾实体的过程。基于该假设,不同模型定义了不同类型的翻译方法和相应的分值评估函数,通过衡量模型翻译头实体的输出结果与真实尾实体之间的相似性,并最小化损失函数学习三元组中各实体关系嵌入,使得正确三元组有取得更大分值。

近年来神经网络的出色表现也为知识图谱链路预测的发展注入了新的活力。ConvE 模型<sup>[44]</sup>对已知的头/尾实体、关系嵌入进行重构并级联拼接得到一个二维矩阵,然后经过二维卷积运算和全连接映射直接得到预测的尾/头实体,通过与目标实体嵌入对比,最小化二维交叉熵损失。ConvKB 模型<sup>[45]</sup>是将每个三元组表示成一个只有三列的矩阵,每列分别与三元组中元素对应,然后输入到卷积神经网络得到一组特征图,再通过特征图拼接得到一个该三元组对应特征向量,最后经过单层全连接神经网络得到该三元组正确性评估值。

### 1.2.2 基于多元关系的链路预测方法

现实知识图谱中除了二元关系事实,同样普遍存在大量多元关系事实,在一些文献中也称为超关系陈述。传统链路预测方法在处理时,通常先将其拆解并转化为二元关系,然后再建模和预测。然而,这种处理方式存在明显的缺陷:一是拆解可能导致部分语义结构信息丢失,如每个多元关系事实仅提取其中一个三元组,而忽略了剩余的实体及关系;二是在转化过程中为了构造三元组,可能设计一些虚拟实体或关系,如一个多元关系事实中将其中的一个三元组整体视作一个实体与其他关系、实体新组成一个三元组,这些虚拟实体关系由于本身不属于知识图谱所以在一些需要构建负样本的方法模型中将对实际模型的训练带来较大干扰,增大模型训练的困难,也增加了模型的计算开销。

鉴于传统链路预测方法在多元关系链路预测问题中的不足,近年提出了针对多元关系事实的链路预测方法,逐渐克服了传统方法的不足。其中最大的区别在于不再需要对多元关系进行拆解或者转化,而是直接在多元关系事实上进行编码建模以及链路预测,直觉上,这种方式显然能更好更完整的提取和利用关系事实中的语义结构信息,使得链路预测的结果更加准确,而在相应实验结果中也有实际的证明。Guan 等人<sup>[3]</sup>最早提出建立在多元关系事实整体上的链路预测方法,提出将一个多元关系事实表示成一组角色-角色值对,即关系实体对的表示形式,认为其可保留关系事实中的全部语义结构信息,同时也提出了相应的链路预测模型 NaLP。实验证明了模型可显示提取多元关系事实中不同关系实体对之间的关联特征。Guan 等人的工作具有开创性,在此基础上后续发展出一些新的模型,但其提出的表示形式存在一定缺陷,比如省略了三元组结构,所有关系实体对属于同等重要性,并且模型忽视了关系事实中不同关系之间的交互。Liu 等人<sup>[7]</sup>同样基于一组关系实体对的表示形式提出了 RAM 模型。RAM 考虑了关系事实的异构性,创新性设计了关系域向量空间以及与实体类型对应的模式矩阵,利用矩阵运算和多线性乘积评估函数训练和预测结果。模型不仅考虑了实体类型与实体之间的关联,也考虑了关系与实体类型之间的关联,具有更高的计算效率和相对更好的预测结果。

除了一组关系实体对表示的多元关系,Rosso 等人<sup>[2]</sup>认为多元关系事实中应该保留三元组结构,提出建立主体三元组与辅助关系实体对表示形式。其依据主要源于现实知识图谱大多基于三元组结构建立,并且作为描述关系事实的基础,三元组在语义语法结构上存在更大合理性。在此基础上提出了 HINGE 模型,实验结果显示其所提出的主体三元组加上辅助关系实体对的表示形式能够帮助提升链路预测效果,从而证明了主体三元组结构的合理性,但 HINGE 在将主体三元组与辅助关系实体对重构成五元组结构的过程中忽视了辅助关系实体对之间的交互信息。

由于知识图谱的图结构特点，图神经网络在方法在知识图谱的各类任务中都得到广泛应用。在多元关系链路预测领域，Galkin 等人<sup>[5]</sup>最早基于图消息传递机制提出了 STARE 模型，基于主体三元组与辅助关系实体对，利用图结构在一个多元关系事实上建立了从实体到关系、从各辅助关系实体对到主体三元组之间的信息流，利用 STARE 作为编码器与其他解码器(如 Transformer)组成的 Encoder-Decoder 架构实现了对关系事实中主体三元组的实体预测。鉴于 STARE 与 Transformer 的编码解码器架构在计算开销上过于昂贵，Yu 和 Yang<sup>[6]</sup>在其基础进行改进，认为 STARE 计算开销过大且作用有限，仅依靠 Transformer 完全能够实现对多元关系事实的特征提取，因此提出了 Hy-Transformer 模型，利用层规范化 LN 和 Dropout 替换 STARE，与 Transformer 建立预测模型。Hy-Transformer 不仅实现了对主体三元组中实体的预测，理论上还可对辅助关系实体对中的实体进行预测，扩展了模型的可用性。上述两个模型忽略了关系事实中关系、实体及实体关系之间的关联类型差异，现实知识图谱本质为异构图结构，即不同实体、关系之间的关联均可为不同类型。Wang 等人<sup>[8]</sup>考虑到多元关系的异构性，基于主体三元组与辅助关系实体对，首次提出用一个简单异构图表示多元关系事实，并构造反映异构信息的一组关联类型变量建立 GRAN 模型。模型利用构造的关联类型变量对 Transformer 中自注意力模块重新设计，并通过 Paddle-Paddle 平台进行了实现。但人为构造的关联类型变量本身不属于关系事实，在固定了关系事实结构的同时也限制了方法的通用性。

除了上述方法，近来又提出一些基于张量分解、几何空间等新方法，在利用多元关系事实信息以及提升效率等方面有所改进，然而这些方法大多只关注到对实体的预测，无法顾及关系预测。

### 1.3 主要贡献与创新

上述工作表明，知识图谱中面向多元关系链路预测任务的主要难点在于深入挖掘多元关系事实中的复杂信息，使其接近现实知识图谱中的关系事实，以及提高事实信息的利用效率。然而现有方法中主要存在对于多元关系事实的理解不足、信息利用不充分问题，包括对关系事实的表达忽略了其中异构性特征；对关系事实的构建时存在部分信息丢失；方法存在设计缺陷造成只能预测实体而无法预测关系的问题。本文围绕增强多元关系事实信息表达、充分挖掘利用事实内部信息这两个目标展开研究，主要创新点和贡献如下：

#### 1. 基于多元关系异构图表示的图变换网络预测方法

针对忽略多元关系事实的异构性以及在关系事实的表达构建过程中存在的信息丢失问题，本文通过分析和理解多元关系事实，提出建立了多元关系事实的异构

图表示形式,在已有关系实体对内部关联、主体三元组与辅助关系实体对之间关联的基础上增加辅助关系实体对之间关联。同时引入图变换神经网络,通过仅利用关系事实本身信息,设计了在大规模图数据集上的图变换网络链路预测方法,利用图变换网络自动生成元路径过程实现对实体、关系的语义结构及其之间异构交互的特征提取与聚合,从而实现对实体和关系的预测。

## 2. 基于轻量化聚合的增强图变换链路预测方法

针对图变换网络受不同关系事实长度变化影响,在不同网络层存在的信息丢失问题,本文通过定性分析与实验设计对其进行了验证,并结合当前图变换网络链路预测模型复杂度较高的情况,提出了一种基于跨通道图最大池化的轻量化聚合方法,认为在不同网络层通道的同一位置,特征值的最大值对应提升模型预测结果的关键特征,从而在将模型各层网络按通道对齐后,利用图最大池化提取分散在各层上的关键特征,实现各层元路径图上关键特征信息聚合,补全了缺失信息。

3. 为了验证在 1 中所提出多元关系事实的异构图表示形式的合理性,作为对照,在其基础上分别建立了两种变体表示形式,一种不含辅助关系实体对之间关联,另一种额外增加主体三元组中实体与辅助关系实体对之间关联。通过设计对比实验证明了所提表示形式的合理性。

4. 为了验证在 2 中提出基于跨通道图最大池化聚合方法认为的不同网络层上关键特征值与最大特征值相对应的合理性,作为对照,分别构建了两种近似平均池化的轻量化聚合方法,一种采用轻量化 LSTM 聚合,一种采用相邻层顺序卷积聚合。通过设计对比实验证明了模型输出特征图上的关键特征对应各网络层中最大特征值的合理性。

5. 针对 1、2 中所提出方法相比于其他现有方法的竞争力,分别设计了对比实验,与基于一组关系实体对表示的主流方法、基于主体三元组与辅助关系实体对表示的主流方法以及部分其他方法进行了对比,实验结果证明了本文方法在部分指标上具有明显优势。

6. 基于前面提出的两个预测模型,结合社交网络中公众人物信息查询的需求,本文设计了一个面向公众人物信息查询的多元关系链路预测系统,提供实体关系预测、数据管理和历史查询功能,并将预测效果在系统中进行可视化展示。

## 1.4 论文结构安排

论文的主要内容包括六个章节,具体内容安排如下:

第一章,绪论。结合实际需求,介绍了本文的研究方向、研究背景以及研究意义。之后,介绍了链路预测领域的国内外研究现状,分别介绍了传统二元关系的链

路预测方法和基于多元关系的链路预测方法，并对后者的发展历程进行了重点介绍。最后，对本文的主要贡献与创新点进行概括并介绍了本文的组织结构。

第二章，相关技术与理论基础。本章对多元关系链路预测任务中所涉及到的知识图谱链路预测、异构图神经网络与图卷积神经网络相关理论，大规模图的图采样技术、池化技术进行了详细介绍，再分类介绍了现有面向多元关系的链路预测方法，最后介绍了本文实验所使用的公开数据集。

第三章，基于多元关系异构图表示的图变换网络预测方法。首先分析了现有多元关系的两种常见表示形式的优缺点，并对基于一组关系实体对表示的方法中存在的不足进行了介绍，然后详细阐述本章所提出基于多元关系异构图表示的图变换网络预测方法的整体设计原理和各实现模块。最后，通过设计与变体表示形式的对照实验以及与基于一组关系实体对表示的主流方法的对比实验分别验证所提表示形式与预测方法的合理性与有效性。

第四章，基于轻量化聚合的增强图变换链路预测方法。首先对第三章引入的图变换网络进行了分析，指出其受不同关系事实长度影响导致的信息丢失问题，其次分析了常用神经网络聚合方法、基于主体三元组与辅助关系实体对表示的预测方法的优缺点。然后，详细介绍了本章提出的基于轻量化聚合的增强图变换链路预测方法的设计与实现思路。最后，通过设计多组对比实验分别验证了信息丢失问题的存在、聚合方法的合理性与有效性以及在多元关系预测问题上的有效性。

第五章，面向公众人物信息查询的多元关系事实链路预测系统。首先介绍了系统的需求分析和总体设计，再具体介绍了系统中存在的各功能模块，最后对系统功能实现界面进行了展示。

第六章，总结与展望。首先总结了本文的主要工作和创新点，然后针对现有工作的可改进之处对后续的研究方向进行了规划，提出进一步的改进思路。

## 1.5 本章小结

本章首先介绍了面向多元关系链路预测的研究背景与意义，然后对近些年来链路预测方法的国内外发展现状进行了总结分析，并归纳现有方法中存在的一些问题，同时引出了本文的主要研究内容与创新点，并在最后介绍了论文组织结构。

## 第二章 相关技术与理论基础

### 2.1 引言

多元关系的链路预测区别于传统的二元关系链路预测，其针对的关系事实可包含更多实体和关系，使得事实的长度具有任意性，语义结构更加复杂，也更接近现实知识图谱中的关系事实。当前多元关系的链路预测方法可大致区分为传统卷积神经网络、图神经网络及其他方法。结合本文所提方法，本章将主要介绍知识图谱多元关系链路预测、卷积神经网络与异构图神经网络相关理论，以及所涉及的图池化、图采样技术，然后概括了面向多元关系链路预测任务的现有方法，最后对常用数据集进行了介绍。

### 2.2 相关理论

#### 2.2.1 知识图谱多元关系链路预测

知识图谱是一种揭示实体与实体之间关系的结构化语义网络，其具有图的基本结构，通过将图中的节点、边与现实中的对象及其之间相互关联建立映射关系，从而可实现对现实世界中事物、事物的描述以及事物之间相互关联进行形式化地描述<sup>[52]</sup>。

知识图谱的架构主要包括信息抽取、知识融合、知识加工和知识更新<sup>[55]</sup>。信息抽取是从外部数据源中抽取出目标实体、关系以及属性等知识，知识融合是对抽取知识的进一步处理，包括消除歧义、建立对应关系以及知识合并等等，知识加工是利用知识做一些更为复杂的计算，如知识推理、本体抽取、知识质量评估，知识更新建立在知识加工的结果上，如利用知识推理的结果更新和完善知识图谱。构建一个高质量的知识图谱涉及到相应的关键技术，包括知识抽取、知识表示、知识合并、本体构建、知识推理等。

当前知识图谱已在学术与工业等领域得到广泛重视。2012 年 Google 公司首次提出知识图谱概念，而随后在 2016 年宣布完成的 Google KG 成为了当年最重要且最受关注人工智能项目之一。许多著名知识图谱，包括 WikiData<sup>[56]</sup>、DBPedia<sup>[57]</sup>、ConceptNet<sup>[59]</sup>、YAGO<sup>[58]</sup>、Satori 等，已成为许多应用及行业研究的重要知识数据库，并为基于知识图谱的上游各类应用领域，包括智能搜索<sup>[60-62]</sup>、知识问答<sup>[63-66]</sup>、智能推荐<sup>[67-69]</sup>等提供了基础支撑。



知识图谱链路预测属于知识推理中的一类研究问题。由于知识图谱主要基于 RDF 类型三元组结构来描述事物之间的事实关系,即由两个实体和一个关系组成的(头实体,关系,尾实体)或者(实体,属性,属性值)作为基础构成<sup>[49]</sup>,因此,链路预测是利用知识图谱中已有的实体和关系,通过模型训练学习,预测给定实体或者实体之间的关系,从而帮助更好地理解知识图谱中的实体与关系。

当前对链路预测问题的研究主要集中在传统二元关系。主要方法包括基于规则的方法、基于概率模型的方法和基于深度学习的方法等,其中利用知识图谱基本组成元素并将其映射到低维向量空间得到知识图谱嵌入,再通过嵌入实现图谱推理的方法在链路预测问题中得到了大量研究。知识图谱嵌入是一种知识图谱表示学习方法,不同于符号表示方法,其核心思想是将知识图谱中的关系、实体映射到一个连续向量空间,从而大大简化计算工作,并能够完全适用于知识图谱上的全部任务,包括链路预测、关系实体分类、关系抽取等,因此该方法一经提出便受到了广大研究者的关注,并不断提出新的嵌入构造方法<sup>[33-34]</sup>,同时也在相关横向领域上利用嵌入方法进行了探索<sup>[46-47]</sup>。

基于嵌入的传统二元关系链路预测方法中有两个经典模型,包括 TransE<sup>[70]</sup>和 RESCAL<sup>[71]</sup>模型。TransE 模型是受到 Word2vec<sup>[72]</sup>中平移特性的启发,提出的第一个基于几何空间的预测模型,其核心思想是将三元组中关系视为头尾实体之间的映射,其中任意一个实体都能通过关系映射得到另一个实体嵌入,模型公式为 $\| \mathbf{h} + \mathbf{r} - \mathbf{t} \|$ ,所基于的假设是一个三元组成立的条件必须是头尾实体与关系之间存在对应关系,使得头实体通过关系映射得到的向量与尾实体嵌入基本一致,模型采用了距离损失函数 L1 和 L2 正则。TransE 模型建立了三元组内部显性的映射关系,在当时具有相当先进性,但实际只能较好的支持一对一的关系,对于普遍存在的一对多、多对一、多对多关系能力有限。针对 TransE 的局限性,后续又提出了许多改进模型,如 TransH、TransD、TransR<sup>[53]</sup>等,使模型适用于更加复杂的情形。RESCAL 模型的核心思想来自三阶张量分解,即高阶张量天然可分解得到低维矩阵和向量。从建模角度,利用关系域自身具有高阶和稀疏特性,可将关系域表示为高阶张量,再通过张量分解得到对应的具体关系和实体。RESCAL 将关系域建模为三阶张量,其中的每个二维平面则对应一种关系,进而可继续分解成一组实体嵌入和一个表示实体之间关联的二维邻接矩阵,从而建立起三元组中关系实体之间的对应关系。该方法相比于 TransE 同样具有广泛应用前景,但由于模型建立在高阶张量基础上,并且其阶数直接受到关系数量影响,因此在大规模知识图谱问题中,受到关系实体数量巨大的影响,会降低对实体、关系的表达准确性,使得后续的任务,如链路预测中对关系、实体的预测结果准确性被降低。

基于嵌入的二元关系链路预测方法可分为基于张量分解的预测模型、基于几何空间的预测模型和基于深度学习的预测模型<sup>[1]</sup>，而知识图谱中除了二元关系外，多元关系也同样普遍存在，并在最近受到越来越重视。与二元关系的固定组成结构不同，多元关系的组成更具有灵活性，其中包含至少两个以上的实体和至少一个关系。因此可将二元关系看作多元关系的一种特例，关系实体之间的关联是多元关系中的基本内在联系，根据多元关系中关系实体对之间的关联类型，又可大致将其划分为一组关系实体对表示的多元关系、主体三元组与辅助关系实体对表示的多元关系、整体不可拆分的多元关系。需要注意的是多元关系的整体不可拆分是指无法将其理解为其其他两种表示形式。

由于多元关系的实体关系数量不受限制，实体关系之间的关联交互也不同于二元关系中的简单映射关系，因此，多元关系的链路预测更加具有难度。对多元关系中的任一潜在待测实体或关系，其与多元关系中的其他已知实体、关系、关系实体对等均可能存在不同程度的关联交互，这种交互的方向可以不同、交互的程度可以不同、交互影响的关系实体数量也可以不同，且同一待测目标在不同多元关系中的关联交互也存在变化，从而大大增加了对其的预测难度。目前针对多元关系的链路预测方法可大致分为基于张量分解<sup>[13]</sup>、路径查找<sup>[15]</sup>、神经网络<sup>[2-4,10,14]</sup>与图神经网络<sup>[5,6,8]</sup>以及几何空间<sup>[7,12]</sup>的方法，这些方法具体将在 2.3 节展开介绍。

### 2.2.2 多元关系链路预测任务中的卷积神经网络

卷积神经网络<sup>[41]</sup>因其具有的强大建模能力而在图像处理、自然语言处理等多个领域得到广泛应用。不同于传统神经网络，卷积神经网络 CNN 在其基础上建立了三种具有不同功能的网络层：卷积层、池化层与全连接层，分别对应其中的卷积操作、池化操作和全连接操作。CNN 的整体网络架构实际通过迭代交替堆叠的卷积层和池化层来构建架构主体，随着网络层数加深，卷积核深度也逐渐增加，从而实现输入图的特征自动提取与聚合，然后通过在网络顶层拼接一个或多个全连接层输出最终的网络预测结果。

卷积神经网络要求处理的目标数据符合欧氏空间，这样数据具有平移不变性，从而支持在输入数据空间定义全局共享卷积核，建立卷积神经网络。文本数据符合欧氏空间，在利用卷积神经网络处理多元关系链路预测问题的方法<sup>[2-4]</sup>中，通过将关系事实中的所有关系与实体特征向量对齐拼接，建立了具有平移不变性的关系事实特征矩阵，即对特征矩阵上任一元素为中心，在采用同一种方式时可获取相同尺寸的局部结构，从而可通过学习作用在每个元素上的共享卷积核来建立局部连接，学习到其中隐藏层表示。然后将关系事实特征矩阵作为卷积神经网络的输入图，

利用卷积与池化操作，分别提取聚合关系实体对内部关联特征以及关系实体对之间的关联特征，最后通过全连接层输出该关系事实的对应向量表示用于后续的预测评估。

卷积神经网络中的两个主要操作是卷积操作和池化操作。卷积操作是利用卷积核，通过滑动窗口的方式，在每个通道上将核上的权值与特征图层上对应位置的特征值进行相乘求和，再对所有通道的输出结果加权求和，得到对应新的特征值。由于卷积核的权值共享特性，同一卷积核得到的输出特征图对应输入特征图层的一种类型特征，因此在需要提取多种类型的特征时，大多采用一组卷积核进行操作，增加了计算的开销。在多元关系链路预测问题中，现有的方法如 HINGE<sup>[2]</sup>和 NaLP<sup>[3]</sup>同样采用了一组卷积核，但将其通道数为设置 1。池化操作通常紧随卷积操作之后进行，主要对卷积输出的特征图层通过类似下采样的方式，对特征进行聚合，从而降低特征维度，进而降低网络的计算复杂度，例如从图像的角度而言，池化可以降低图片的分辨率。常见的池化包括平均池化、最大池化与最小池化，即在卷积核大小区域内，分别对区域内部的特征值计算平均值、选取最大特征值和最小特征值作为该选定区域内所有特征值的替代，从而实现特征减少。

### 2.2.3 面向知识图谱的异构图神经网络

近年来，关于图的深度学习在研究领域蓬勃发展，其中图神经网络模型<sup>[48]</sup>是最为成功的学习架构之一，其通过将循环神经网络拓展到图结构，直接在图数据上建立神经网络，通过聚合图节点邻域信息和节点当前状态信息来迭代更新节点表示，使其学习到图拓扑结构等特征，从而实现对事物间复杂相互作用关系建模。目前，关于图神经网络的研究已经渗透到许多其他领域，包括推荐系统<sup>[26]</sup>、计算机视觉<sup>[36]</sup>、自然语言处理<sup>[33-34]</sup>、知识图谱<sup>[5,8,38]</sup>等等，称为各领域争相研究的热点问题。

由于知识图谱本质是一类复杂多样的异构图，因此图神经网络在知识图谱相关问题中提出的方法或模型，一定程度上属于异构图神经网络。异构图由不同类型的节点以及节点之间的关联（称为边）构成，在语义内容和结构上都具有异构性，也叫异质性。例如一个简单社交网络中同时存在用户、评论、发布文本或图片等类型的节点，以及用户-评论、用户-发布文本、用户-发布图片、发布文本-评论、发布图片-评论等多种关系，节点之间的语义关系具有多样性，并且节点和边所对应的属性特征可以处于不同特征空间。目前关于异构图的研究存在几个难点<sup>[23]</sup>，包括 1) 图节点关联结构的异构性，即对不同的节点，其邻域内连接的节点类型和数量都可以不同；2) 不同节点可以存在多模态语义信息，即节点语义内容可以来自文本、图像等半结构与无结构数据；3) 中心节点邻域内存在不同相邻节点对中心

节点的影响差异性。针对以上问题,研究者提出了许多相关方法,如 HetGNN<sup>[23]</sup>、HAN<sup>[24]</sup>、HGT<sup>[22]</sup>、GTN<sup>[20]</sup>、HGSL<sup>[35]</sup>,分别从消息传递机制与邻域聚合的角度、图注意力机制的角度以及元路径图生成的角度结合图神经网络提出相应的解决方案。元路径是一种由不同类型的节点及其之间关联所组成的一条串联路径,属于异构图的基本结构,一些方法认为通过在元路径上建立消息传递与邻域聚合机制,从而得到的节点表示将重新具有同质性,即能够因此将原本的异构图转化为同构图,从而结合传统基于同构图的图神经网络等其他方法应用于各类上层应用领域,如 HGT 是基于 Transformer 在异构图上的适应性改进,GTN 是通过堆叠多个图变换层来建立元路径生成过程,然后利用图卷积运算提取各层中的不同的特征信息。在知识图谱的多元关系链路预测问题中,目前很少有方法涉及图谱的异构性并建立基于异构图的预测方法,最近一篇文献中提出的 GRAN<sup>[8]</sup>模型,是通过在多元关系中人为设置四种特定的链接类型变量,引入 Transformer 的自注意力机制模块,从而构建基于异构图的多元关系链路预测模型,但这种人工变量的引入与对 Transformer 的改造方法受限于目标数据集本身特征以及作者本人对多元关系结构的理解,从而一定程度限制了方法在其他领域的通用性。

## 2.3 相关技术

### 2.3.1 图采样

现实知识图谱通常规模巨大,由于节点之间的相互关联,在图上直接使用图神经网络进行多层连接后节点的计算量和中间变量的规模均呈现爆炸式的增长,常常超出硬件设备的存储上限。因此,如何缩小每次运算中输入的图的规模,使其能够成功运行图神经网络模型,成为了具有现实意义的研究重点。图采样便是解决问题的一种处理技术,其与一般采样的目标一致,即从候选图数据集中抽取出一组由目标节点所构成的子图数据集,然后进行下一步的处理和分析,不同之处在于,图上采样还需要符合节点之间的关联。图采样能够有效帮助一些常规图神经网络适应大规模图数据,并缩短获得有用结果的计算时间,还能帮助研究人员更好地理解图结构。

目前图采样已经在图像、自然语言处理、知识图谱等多个领域得到应用,在知识图谱链路预测问题中,图采样技术的具体应用包括改善图神经网络模型的可拓展性,提高知识图谱的可搜索性,结构可解释性和推理性。图采样技术按照采样粒度又可以分为邻域采样、层采样和子图采样。Hamilton 等人<sup>[16]</sup>提出图采样聚合网络模型 GraphSAGE,不同于其他模型建立节点表示时需要考虑所以邻域节点,

GraphSAGE 可对邻域节点随机采样,在大规模图数据集中减少每次输入的节点数,使得在大规模图数据集上搭建的图神经网络成为可能。由于每个节点都需要进行邻域采样,而一个节点可以出现在多个节点的邻域中,因此邻域采样可能出现节点的重复采样问题,这些重复采样大大增加了模型的计算量,使得模型效率低下。另一方面, FastGCN<sup>[19]</sup>中建立了一种面向网络层的重要性采样方法,利用节点度数计算相应的采样概率,并根据概率在每个网络层上采样固定数量的节点,然后利用这些采样节点构建新的更小的图邻接矩阵用于网络计算,从而降低整体模型的计算开销。然而这种利用节点度数计算的采样概率在不同层之间具有独立性,相邻层上的节点因此可能不存在关联,使得各网络层上构建的图邻接矩阵特别稀疏,即存在图结构信息丢失的问题,从而大大降低模型的性能。针对两种经典采样方法中存在的问题, Zhou 等人提出了 LADIES<sup>[18]</sup>采样模型,结合了邻域采样与层采样的优点并在层采样基础上融合了重要性,从而提升了采样效率。此外,由于上面几种采样模型主要用于相对小规模图上, Zeng<sup>[17]</sup>等人基于深度 GCN,利用归纳学习在大规模图上通过子图采样方法建立了 GraphSAINT 模型,从而克服深度神经网络中邻域爆炸问题,加速模型训练过程。

### 2.3.2 图池化

图池化可以视为普通神经网络中的池化在图神经网络的推广。池化作为特征数据降维、特征聚合的一种有效方法在神经网络中得到大量运用,如卷积神经网络中的池化层,通过每次对卷积核视野范围的特征值采用池化操作,计算得到范围内所有特征值的等效替代,从而实现在指定维度上的特征降维。类似的,图池化是图神经网络架构中的重要组成,图神经网络通常可视为由学习节点特征的特征变化层与减少节点数量的池化层交替构成,同时,图池化层又可视为一种独立操作,其输出的结果是对应图级别的完整特征图层<sup>[40]</sup>。

目前图池化的研究进展已经出现类似于消息传递机制的可学习运算机制,在训练过程中实现动态自适应具体任务的最优池化计算,其中包括 DiffPool<sup>[42]</sup>这样的聚类方法,以及大量的学习保留部分节点而丢弃其他节点的方法。此外,还有许多将整个图简化为单一向量表示的全局池化方法。

一种针对图池化方法的简洁分类是将其划分为扁平池化和分层池化<sup>[43]</sup>。扁平池化又被视为图的读出操作,是在图上通过单步运算,如对整个图上所有节点嵌入直接计算平均值或者最大值,从而一次得到图级别的输出表示。分层池化则是通过迭代每次生成一个更小的新图,从而逐步将输入图简化,池化过程保持了输入图的层次结构特征。分层池化在具体方法上又可分为节点聚类池化、节点丢弃池化和其

他池化方法，其中前两者的区别在于节点聚类池化是通过对上一层的图上节点聚类，然后将生成簇作为构建下一层图上的新的节点，则这种新生成的节点本身是原图上一组节点的特征聚合，而非原图上存在的节点；节点丢弃池化则是从原图的所有节点中每次筛选出一个节点子集构成新的简化图层，其中的每个节点仍属于原图上的节点。因此，相比之下，节点聚类池化需要更大的计算开销但也基本保全了输入图上的所有信息，而节点丢弃池化的计算效率更高也更适应于大规模图问题，但相对而言其不可避免存在一定的信息损失。需要注意的是，扁平池化方法中所采用的平均值计算、最大最小值计算同样可适用于分层池化中每层上简化图的构建。

### 2.3.3 面向多元关系的链路预测

多元关系是由两个及以上实体、一个以上关系构成的复杂关系。相比于传统二元关系，多元关系事实能够表达更复杂的信息，贴近现实中的语言表达，因此能够更好地促进对语言文本的理解，满足用户相关需求。

多元关系链路预测作为知识图谱链路预测问题的一个重要分支领域，其研究发展的时间并不长。早期的研究中通常是将多元关系分解为二元关系，然后利用二元关系的方法来完成预测任务。如 2016 年最早提出的 **m-TransH**<sup>[9]</sup>模型，通过对多元关系事实重构转化为二元关系对应的三元组来实现多元关系的链路预测；之后在 2018 年又提出了 **RAE**<sup>[11]</sup>模型，是针对前面 **m-TransH** 模型只考虑了三元组内部关联的情况下，增加了关系事实中各个三元组中的关系之间的交互关联信息，减少了信息丢失。但这些方法从思想上仍然与传统的二元关系链路预测大致相同，无法解决其在多元关系问题存在信息丢失与事实歧义问题。

从 2019 年的 Guan<sup>[3]</sup>等人首次提出的 **NaLP** 模型开始，后续的研究成功将完整的多元关系事实作为模型的直接处理对象，并在此基础上建立了直接面向完整多元关系的链路预测模型。目前，根据模型所采用的具体方法，可划分为基于张量分解<sup>[13]</sup>、路径查找<sup>[15]</sup>、神经网络<sup>[2-4,10,14]</sup>与图神经网络<sup>[5,6,8]</sup>以及几何空间<sup>[7,12]</sup>等方法，其中在基于神经网络的方法中包括了 **NaLP** 系列（**NaLP**、**tNaLP** 与 **tNaLP<sup>+</sup>**<sup>[11]</sup>）和 **HINGE** 在内的经典模型，通过组合 **CNN** 和 **FCN** 来实现关系事实中语义交互特征提取和聚合，建立了关系事实的整体嵌入表示，然后通过对事实嵌入进行评估来预测其中的潜在目标；基于图神经网络的方法中目前只有 **STARE**<sup>[5]</sup>和 **GRAN**<sup>[8]</sup>这两种模型，通过将多元关系理解为主体三元组与辅助关系实体对的基本形式，其中 **STARE** 模型采用图消息传递机制和邻域聚合构建编码器，将辅助关系实体对中的信息流向主体三元组并与之聚合，建立了针对主体三元组中头尾实体的嵌入表示，实现对主体三元组中主实体的预测，整体预测模型采用编码-解码器架构，**STARE**

对应的编码器与 Transformer 对应的解码器组合，作为当时的 SOTA，得到在主实体预测上较准确的结果，但该模型基于同构图的基础逻辑，没有考虑到多元关系事实本身关系实体的异构性，并且方法设计中存在一定缺陷，只能实现对主实体的预测，忽略了关系预测以及辅助关系实体对中的关系、实体预测，同时，STARE 与 Transformer 的组合模型复杂度过高，计算开销偏大。GRAN 模型在分析中认为 Transformer 足够很好地实现对文本语义信息的提取，因此可直接用 Transformer 实现多元关系事实特征的提取和融合，同时考虑到多元关系的异构性，为多元关系事实建立了的异构图表示形式，即对关系事实中主体三元组的主关系与各辅助关系实体对的辅助关系建立连接，与主体三元组内部关系与头尾实体之间连接、各辅助关系实体对内部关系实体之间连边共同构成异构图结构。为了克服 Transformer 基于同构图建立而非异构图的特点，GRAN 针对异构图中的不同连边人为设置了四种对应连接类型变量，与其所构建的异构图中连接节点的边的类型相对应，并将链接类型向量融合到 Transformer 的自注意力模块中，使其对异构图具有兼容性。鉴于 GRAN 模型实际引入了外部信息，而非纯粹利用关系事实本身信息的特点，与其他方法存在明显区别，因此本文不考虑将其作为所提方法的对照模型。基于张量分解与基于几何空间的方法本质都是从借用数学高维几何空间的思路对多元关系事实建模，其中的方法大多是将多元关系事实作为一个难以分解的复杂抽象关系，通过高维张量分解或者是向量空间的极大无关组来自适应学习其中的关系实体嵌入，方法在计算效率是具有一定优势，但由于这些方法中大多将关系作为生成实体嵌入中的类似映射函数，因此方法的实际预测对象都只是对实体进行预测，无法预测关系。路径查找的方法目前只出现在一篇针对三元关系的知识图谱链路预测文献中，还未能就复杂的混合多元关系进行拓展。

## 2.4 主要数据集介绍

研究多元关系链路预测问题的常用数据集包括 WikiPeople<sup>[3]</sup>、JF17K<sup>[9]</sup>、WD50K<sup>[5]</sup>及其派生数据集，如 WikiPeople-4（仅由 Wikipeople 中的 4 元关系组成的子数据集）、JF17K-4、WD50K\_100 等。考虑到 WikiPeople 数据集中多元关系事实的比例不到 3%，这里主要介绍 JF17K 和 WD50K 数据集。

### 2.4.1 JF17K

JF17K 数据集取自 Freebase，其中共包含 100947 条关系事实实例，28645 个不同实体与 322 个不同关系，超过二元的多元关系实例占比达到 45.9%。在该数据集中，多元关系通过一个复合关系/属性及其包含的实体来进行定义，并且属于该

复合关系下的多元关系事实应包含完整的相应属性值（即实体），如数据集中的“soccer.football\_player\_match\_participation”表示一个复合关系，则所有包含这个关系的多元关系事实都需要包含至少两个实体。该数据集的最大元数为 6，即一个多元关系事实中最多包含 6 个实体。在最近的少数论文中也能见到由 JF17K 派生的子数据集，如 JF17K-3、JF17K-4，分别指从 JF17K 中筛选仅包含 3 元或 4 元的关系事实所构成的子数据集。通常这类子数据集被方便用于一些特定问题的研究实验，如研究三元关系事实中的链路预测方法<sup>[15]</sup>。

### 2.4.2 WD50K

WD50K 抽取自 Wikidata，其中共包含 236507 个关系事实实例，47156 个不同实体和 532 个不同关系，而超过二元的多元关系实例占比达到 13.6%。由于其构建时利用数据集 FB15K-237 中的实体对 Wikidata 直接映射，使得获得的关系事实均由相应的实体关系代号（token）组成，内容简洁清楚但字母语义丢失。不同于其他数据集，WD50K 数据集中关系事实的实体分布更宽广，并随机划分组成训练测试和验证集。WD50K\_100 是 WD50K 中筛选多元关系实例得到的子数据集，其中多元关系事实占比达到 100%，共有 31314 条多元关系实例，18792 个实体和 279 个关系，按照 7:1:1 随机划分训练集、验证集和测试集。

## 2.5 本章小结

本章首先介绍了知识图谱多元关系链路预测、多元关系链路预测问题中的卷积神经网络与知识图谱中异构图神经网络相关理论，然后结合本文中所提出方法，介绍了方法涉及到图采样技术、图池化技术，然后分别阐述了当前多元关系链路预测相关研究重点的方法类型以及对应相关模型，最后对问题研究中的常用数据集进行了介绍，为后续第三章和第四章的内容提供了理论和技术基础。



## 第三章 基于多元关系异构图表示的图变换网络预测方法研究

当前对于多元关系的理解大致可分为基于一组关系实体对的表示方法和基于主体三元组与辅助关系实体对的表示方法。前者从更细粒度拆解关系事实并认为各关系实体对之间具有同等重要性，但这类方法在拆解过程中容易出现语义结构信息丢失，如忽视了关系事实中目标内容的主体性而出现事实歧义，并且表示方法忽视了事实中异构性特征；后者虽然突出了主体内容，但大多数方法中仍然忽视了关系事实的异构性。本章首先对多元关系的两种典型表示方式所存在的信息丢失问题进行了介绍，并重点分析了基于一组关系实体对表示的多元关系链路预测方法的优缺点，然后详细阐述了基于多元关系异构图表示的图变换网络链路预测方法的建模思路和模型架构，最后在公开数据集 JF17K 上构建了对比分析实验，结果表明本章方法在 MRR 指标上，对比 2021 年 tNaLP<sup>+</sup>模型在混合关系预测提升了 0.52%，混合实体预测提升了 1.36%。

### 3.1 引言

与固定了实体与关系数量的二元关系不同，多元关系通常不会限制其具体数量，除了可包含二元关系外，还可以在此基础上增加任意数量的关系实体。因此，多元关系可以表达更复杂的关系事实，更贴近现实中文本。传统的链路预测方法无法直接处理这样的复杂关系事实，只能通过人为构造虚拟实体关系将其分解成一组三元组或者仅截取其中一个三元组用于模型预测，从而造成大量信息丢失问题，降低了预测准确性。

当前针对完整多元关系事实建立的多元关系链路预测模型，在对多元关系事实的理解上主要有两种表示方法，一类基于一组关系实体对表示，另一类基于主体三元组和辅助关系实体对表示。以一段文本为例，“瑞典皇家科学院宣布，将 2021 年诺贝尔化学奖授予德国科学家本杰明·利斯特（Benjamin List）和美国科学家戴维·麦克米伦（David MacMillan），以表彰他们‘在不对称有机催化的发展’方面的贡献”，其中的关系有（宣布、授予、表彰、人物、时间、国家），实体有（瑞典皇家科学院、诺贝尔化学奖、本杰明·利斯特、戴维·麦克米伦、美国、德国、2021），则这些实体和关系可构成这样一个多元关系事实：“2021 年瑞典皇家科学院宣布诺贝尔化学奖授予德国本杰明·利斯特和美国戴维·麦克米伦。”如图 3-1 所示。

传统的二元关系链路预测需要对这个关系事实截取信息，如只提取其中一个三元组：（瑞典皇家科学院，宣布授予，诺贝尔化学奖），因此丢失了许多重要信息。

而在针对多元关系事实的两种表示形式下，关系事实可分别表示为图 3-1 中的两种形式。

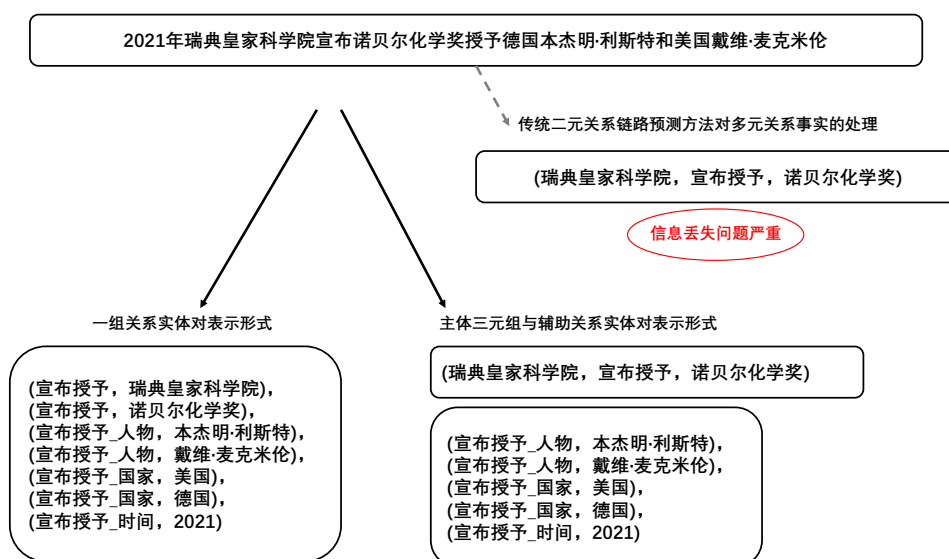


图 3-1 多元关系事实的不同表示形式

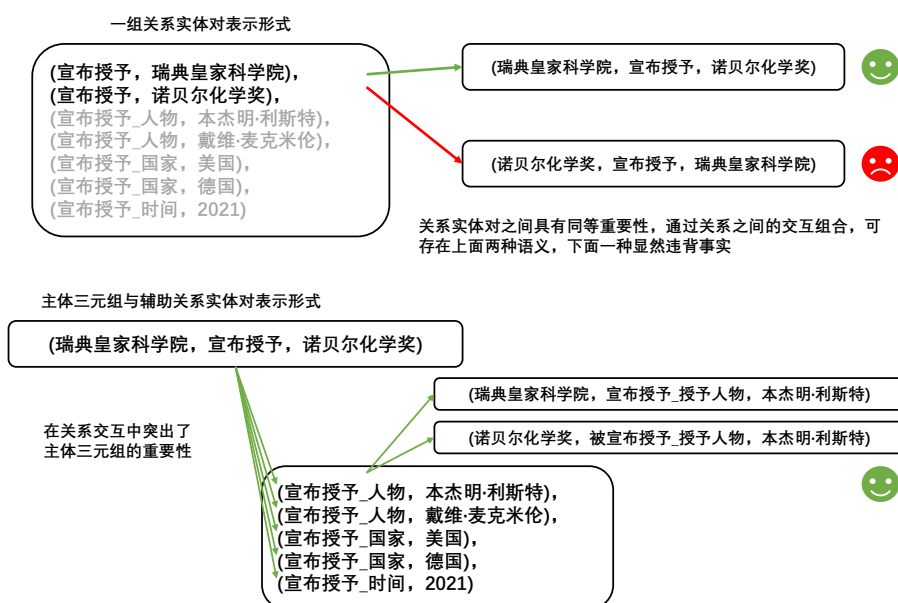


图 3-2 由一组关系实体对表示的多元关系所存在的信息丢失问题及其导致的事实歧义

由于一组关系实体对的表示要求其中各关系实体对具有同等重要性，导致原本关系事实中的主体内容在语义结构上丢失了主体性，一种显而易见的错误在于通过关系之间的交互，其中（宣布授予，瑞典皇家科学院）和（宣布授予，诺贝尔

化学奖)在组合后,既可以表达为“瑞典皇家科学院宣布授予诺贝尔化学奖”,也可以是“诺贝尔化学奖宣布授予瑞典皇家科学院”,显然后者存在事实性错误,其组成的关系事实“2021 年诺贝尔化学奖宣布授予瑞典皇家科学院的德国本杰明·利斯特和美国戴维·麦克米伦”则出现事实歧义的问题,如图 3-2 所示。类似的组合也可以发生在其他任意两个关系实体对之间,因为其具有的同等重要性。而在主体三元组与辅助关系实体对表示下,由于确定了关系事实中的语义主体作为三元组,使得三元组在语义和结构上均具有高于辅助关系实体对的重要性,因此可避免出现上面的事实性错误问题。此外,当前知识图谱大多将三元组结构作为其图谱构建时的基础要素,并利用三元组来存储关系事实的主体关系与实体,因此三元组的实际重要性也明显大于其他关系实体对<sup>[2]</sup>。

然而,在图 3-2 的例子中,两种表示形式实际都没有体现出多元关系事实的异构性特征,如主体三元组中实体与关系之间的关联由于其本身增加了语义结构的重要性,应该与辅助关系实体对中实体与关系之间的关联存在一些差别;同时,关系事实中实体与关系之间的关联、主体三元组与各辅助关系实体对之间的关联、以及各个辅助关系实体对之间的关联也同样应该存在类型差异。因此,像图中这样无论是采用一组关系实体对表示形式还是主体三元组与辅助关系实体对表示形式,实际都至少存在忽略关系事实的异构性特征,造成了信息丢失。通过补充多元关系的异构信息,改进现有的表达形式,使其贴近多元关系事实的原始语义,可认为能够有效提升关系实体预测的准确性。

基于对多元关系的分析理解,针对上述不同表示形式所实际存在的问题,本章提出了多元关系的异构图表示形式,并在现有关系实体对内部关联、主体三元组与辅助关系实体对之间关联的基础上增加了辅助关系实体对之间关联。为充分挖掘多元关系异构图上语义结构信息,提出引入图变换网络,实现同时提取关系事实中的关系、实体异构信息及其之间异构交互并生成元路径图,通过多层网络堆叠挖掘出关系事实深层信息,实现对事实中潜在关系或实体的预测。

### 3.2 基于一组关系实体对表示的相关方法分析

当前基于一组关系实体对表示的主要多元关系链路预测方法包括 NaLP<sup>[3-4]</sup>系列模型、RAM<sup>[7]</sup>模型。神经网络模型 NaLP<sup>[3]</sup>通过提出将多元关系事实表示为一组关系实体对的表示形式,在关系事实所对应的一组关系实体对上直接提取其中的内在交互特征,从而建立预测模型。该方法为事实中关系和实体建立相同长度的初始嵌入,并将每个关系实体对内部关系与实体嵌入级联、关系实体对之间长度对齐拼接构成关系事实对应特征矩阵,然后利用卷积神经网络和全连接神经网络分别

提取关系实体对内部以及不同关系实体对之间的关联交互特征，得到关系事实在特征聚合后对应的一组特征向量，通过假设有效关系事实成立的基础是关系事实对应特征嵌入的任意位置上特征元素值都应足够大，因而训练后各位置上与有效关系事实嵌入相关的最小特征值也应尽可能大，故采用最小池化实现对前面得到的一组特征向量在每个位置上的特征聚合，生成最终关系事实在特征聚合后的对应特征向量并用于后续全连接神经网络模块的预测评估。

对于多元关系事实，NaLP 虽然考虑了每个关系实体对内部以及不同关系实体对之间的交互，但仍然忽略了关系事实的异构信息以及关系事实的内容主体性。针对前者，作者在 NaLP 基础上后续又提出了 tNaLP 和 tNaLP<sup>[1]</sup>模型，试图引入关系类型约束，作用近似于关系异构性，一定程度上补充了事实信息；同时，设计了一种新的负采样方法，采用整体替换关系事实中某个实体关系对而非传统的仅对其中某个关系或者实体进行替换，从而增大了正负样本之间的差异性，一定程度也提高了预测准确度，但这些工作并没有从模型本身实际提升对关系事实的信息挖掘能力，也没有完全解决关系实体异构信息缺失与事实内容的主体结构缺失问题，模型效果提升有限。

最近的 RAM 模型在多元关系事实的关系与实体之间新增了一类称为“角色”的变量，从作用上可理解为关系类型变量，例如在复合关系“在电影中扮演”与三个实体“施瓦辛格”、“T-800”、“终结者 2”之间建立了“演员”、“角色”、“电影名”三个对应“角色”变量，作为对复合关系的内容补充。因此，建立了关系与关系类型为一对多，关系类型与实体为一对一的表达结构。RAM 从关系域出发，为每个关系实际构建了一个潜在向量空间，其中每个关系类型都是组成向量空间的一个基础向量，并通过对应模式矩阵与不同实体嵌入之间建立非线性映射关系。与其他方法不同在于，模型主要从关系及关系类型角度，通过融合相应权值来建立关系之间的语义关联。同时，采样多层嵌入方法将每个实体嵌入表示为一个二维矩阵，增强了不同实体之间的差异性表达，并建立了从实体到关系类型的信息流。然而，RAM 模型在实际上并没有真正考虑关系事实中的异构交互性，只是对关系与实体变量的嵌入维度上进行了区分，并且方法从实验上只验证了在混合实体上的链路预测效果，未能对关系进行验证。

上述方法，除了它们各自在设计中存在的一些问题，还都普遍存在由一组关系实体对表示方法所带来的结构性缺陷问题。针对这些问题，本章提出了基于多元关系异构图表示的图变换网络链路预测方法，该方法具有如下优点：

(1) 建立的异构图表示形式能够兼具多元关系的语义结构信息与异构特征，为模型更充分地挖掘关系事实信息提供基础；

(2) 首次在多元关系预测中引入图变换网络, 利用其元路径以及元路径图生成过程充分提取和聚合异构图上节点特征信息, 同时为实体和关系建立特征嵌入, 在充分挖掘事实信息的同时实现对实体和关系的预测。

### 3.3 基于多元关系异构图表示的图变换网络预测模型

针对多元关系的现有表示方法中所存在的不足, 本章基于主体三元组与辅助关系实体对的表示方法, 提出建立多元关系的异构图表示形式, 并在关系实体对内部关联、主体三元组与辅助关系实体对之间关联的基础上增加辅助关系实体对之间关联; 同时, 为充分挖掘多元关系中关系、实体之间异构交互特征, 引入一种图变换网络, 对异构图中的关联路径实现特征提取和聚合, 从而实现对潜在关系、实体的预测。

#### 3.3.1 总体模型框架

基于多元关系异构图表示的图变换网络链路预测模型总体框架如图 3-3 所示。模型整体流程包括输入子图生成过程、图特征提取与聚合过程以及目标节点预测过程, 其中输入子图生成过程包含 LADIES 采样模块和降噪模块、图特征提取与聚合过程包含图变换网络模块、目标节点预测过程包含 FCN 模块。

模型首先经过输入子图生成过程, 大规模输入图邻接矩阵 $A$ 与其对应的基于多元关系异构图表示中关系实体之间的不同连接类型划分得到的一组邻接矩阵 $A$ 无法直接应用于图变换网络, 因硬件设备的限制以及其存在较低的计算效率。因此, 通过小批量学习, 每次以一个 Batch 内的节点为初始节点, 利用 LADIES 采样模块从图邻接矩阵 $A$ 上围绕初始节点对其多跳邻域内节点循环采样, 得到候选节点集合, 再将其对应到邻接矩阵组 $A$ 中, 得到实际输入的子邻接矩阵组 $A'$ 。然后经过降噪模块来增强子邻接矩阵的鲁棒性, 生成后续图变换网络的输入子图 $\varphi(A')$ 。

在图特征提取与聚合过程, 图变换网络每次从输入子图 $\varphi(A')$ 中自适应筛选出具有相对强关联性的子邻接矩阵作为网络层输入, 与上一层输出的元路径图通过矩阵乘法实现图上元路径自动生成, 得到当前层上的元路径图输出。模型通过迭代 $l$ 层对图上节点及交互特征实现提取与聚合, 输出最后一层上元路径图用于目标节点预测。

最后, 模型的目标节点预测过程通过堆叠两层 FCN 来实现, 节点嵌入与聚合了图特征的元路径图融合, 得到图上所有节点对应图特征聚合的新嵌入, 通过计算目标待测节点与这些节点的相似度及排名, 得到目标节点最终的预测评估分值 $P$ 。

后续小节将主要对前两个过程中的模块展开介绍。

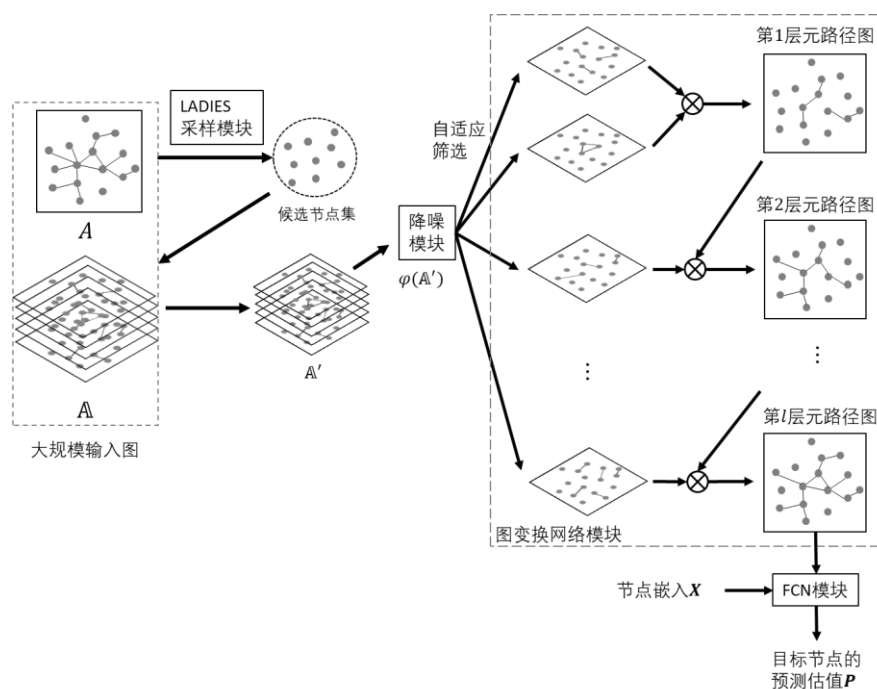


图 3-3 基于多元关系异构图表示的图变换网络链路预测模型架构图

### 3.3.2 输入子图生成过程

#### 3.3.2.1 多元关系的异构图表示

多元关系事实的异构图表示是利用异构图对多元关系中的关系、实体、以及关系实体之间不同类型的连接建立的表示形式。由于图结构的灵活性，相比线性表示其能够较好的建模不同关系实体之间的复杂交互，又有多元关系事实本身内在的异构性，因此异构图表示形式相比其他表示形式更符合多元关系事实信息的表达。以一个多元关系事实  $\mathcal{S}$  为例，基于主体三元组与辅助关系实体对得到形式表达为  $\mathcal{S} = \{(s, r, o), (a_1, v_1), \dots (a_n, v_n)\}$ ，而关系事实  $\mathcal{S}$  映射到异构图上的结构如图 3-4 所示，其中实体（圆表示）和关系（方块表示）都作为图中节点，实体与关系之间的连接作为图中连接节点的边，则根据关系事实中实体关系之间的潜在交互类型，可构成至多六种类型的边，包括主体三元组中头实体  $\leftrightarrow$  主关系、尾实体  $\leftrightarrow$  主关系，主关系  $\leftrightarrow$  辅助关系、辅助关系  $\leftrightarrow$  辅助实体、头实体  $\leftrightarrow$  辅助关系、尾实体  $\leftrightarrow$  辅助关系。然而，考虑到头实体  $\leftrightarrow$  辅助关系、尾实体  $\leftrightarrow$  辅助关系两种连接存在将辅助关系实体对的重要性提升为主体三元组，从而实际上模糊主体三元组的结构特征的潜在负面影响，本章中提出的多元关系异构图表示实际是由关系实体节点与前四种类型的连接所构成的异构图形式，称为  $\mathcal{G}_{h-std}$ 。

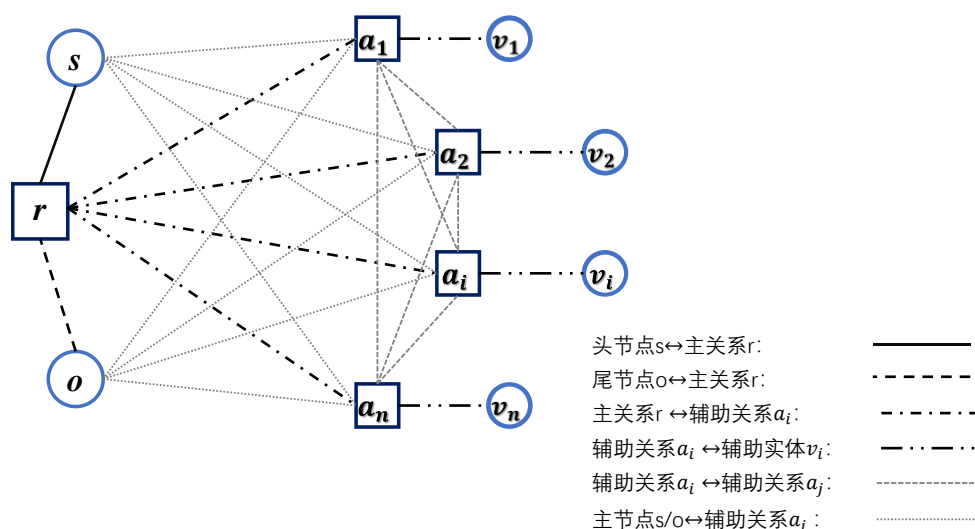


图 3-4 多元关系异图的完全连接表示形式  $G_{h-full}$ ，包含可能存在的全部六种边类型

令异构图表示为  $G_h = (\mathcal{V}, \mathcal{E}, \mathcal{T}_v, \mathcal{T}_e)$ ，其中类型映射表示为  $\tau_v(v) = \mathcal{V} \rightarrow \mathcal{T}_v$ ， $\tau_e(e) = \mathcal{E} \rightarrow \mathcal{T}_e$ 。为验证本章所提出的由关系实体节点与四种类型连接所构成的多元关系异构图表示的合理性，作为对比，相应建立了两种异构图变体表示，即一种删除了辅助关系之间以及头尾节点与辅助关系之间的连接，由关系实体节点与头实体  $\leftrightarrow$  主关系、尾实体  $\leftrightarrow$  主关系，主关系  $\leftrightarrow$  辅助关系三种连接所构成的变体，称为  $G_{h-base}$ ，这种形式可用来对比验证辅助关系实体对之间可能存在的关联及其对潜在目标预测的影响；另一种由关系实体节点与全部六种连接构成的变体，称为  $G_{h-full}$ ，如图 3-4 中所示，可用来对比验证前面提到的主体三元组的重要性。本章后面将分别基于这三种异构图表示进行链路预测的相关对比实验。

### 3.3.2.2 图采样模块

面对大规模图数据集，实体关系的巨大数量使得图神经网络无法直接运用，对此，常见的一类方法是利用图采样每次提取其中一部分节点或者子图输入到模型进行处理，最后将所有输出结果综合。本章中借鉴这一思路，考虑采用 LADIES 图采样模型<sup>[18]</sup>。LADIES 分别借鉴了邻域采样 GraphSage<sup>[16]</sup>与分层采样 FastGCN<sup>[19]</sup>的优缺点，采样一种基于中心重要性的采样方法，既包含了层采样思想，减少了层内各个中心节点之间的重复采样，又包含了邻域采样思想，保留了相邻网络层上采样节点之间的关联性。此外，LADIES 还为候选采样节点增加重要性参数，提高了采样目标的准确性并降低了采样方差。

LADIES 模型的执行过程可表述如下：（1）整体采样过程由中心向邻域拓展，在网络最上层上输入一组节点，则模型将围绕这组节点完成采样过程；（2）在当前网络层上，以上一层采样得到的节点作为新的中心节点，获取其对应在当前层上除自身外的所有邻域内节点，并在这相邻两层网络节点之间构建双线图；（3）对前一步得到的所有邻域节点根据节点在当前层上的度数计算各自的采样概率，并基于采样概率对节点在当前层采样；（4）模型在每层网络上利用采样得到节点构建本层对应子邻接矩阵，并经过行规范化输出，用于后续的训练和推理。

本章在采用 LADIES 模型时做了适应性调整，考虑到多元关系数据集的稀疏性以及多元关系事实中实体关系之间关联的复杂性，因此将当前层上的节点采样范围调整为包含上一层所有中心节点以及及其邻域节点之和，同时在每层上分别设置了对应最大节点采样数量，使得采样过程优先对重要性高、采样概率大的节点。采样过程如图 3-5 所示，

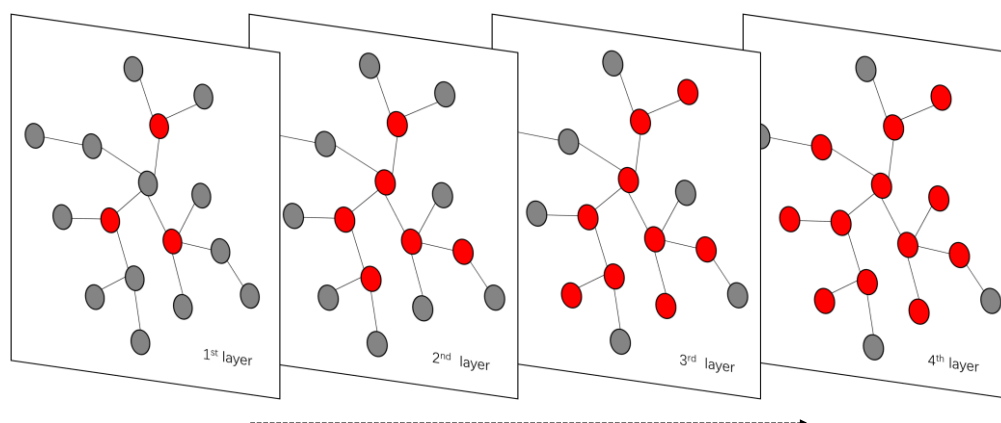


图 3-5 LADIES 模型采样过程示意图

模型采用自顶向下采样，网络层上任意节点采样概率计算公式如下， $(i \in N)$

$$p_i^{(l-1)} = \frac{\|Q^{(l)}P_{*,i}\|_2^2}{\|Q^{(l)}P\|_F^2} \quad (3-1)$$

式中 $Q^{(l)}$ 为第 $l$ 层上采样输出中心节点集合， $Q^{(l)}P \in \mathbb{R}^{|Q^{(l)}| \times N}$ ,  $l = L \dots 1$  对应在第 $l-1$ 层上行采样矩阵，即 $Q^{(l)}$ 关于其在图拉普拉斯矩阵 $P$ 的对应行拼接， $\|Q^{(l)}P\|_F^2$ 反映了中心节点在图上关于所有节点的关联强度。

### 3.3.2.3 降噪模块



实际多元关系数据集中,不同数据集之间、不同多元关系事实之间可能存在语义结构的差异性,即不同关系事实中的实体与关系之间的连接数量、甚至连接类型数量都可能不同。因此,为了增强模型的鲁棒性以及泛化能力,在多元关系异构图表示的基础上,设计增加一个降噪模块,在 LADIES 采样得到的子图(即一组按连接类型划分的子邻接矩阵)输入到图变换网络之前,利用哈达玛积为每次输入的子邻接矩阵乘上一个降噪矩阵,即等效于对子图 $\mathbf{A}'$ 进行如下计算,

$$\varphi(\mathbf{A}') = \text{Hadamard}(\mathbf{A}', \mathbf{z}) \quad (3-2)$$

式中 $\mathbf{z}$ 表示一组具有相同数量和规模的降噪矩阵。从而实现在训练过程中,降低对具体的多元关系事实中可能不存在的关联所产生的影响。

### 3.3.3 图特征提取与聚合过程

图特征提取与聚合过程是利用图变换神经网络实现对异构图上节点与边的特征提取与聚合。图变换神经网络 GTN<sup>[20]</sup>与 Fast-GTN<sup>[21]</sup>是一种建立在异构图上的图神经网络模型,模型底层原理在于异构图中的元路径相关概念以及图卷积神经网络 GCN<sup>[54]</sup>。在异构图中,一条多跳连接的路径被称为元路径,  $v_1 \xrightarrow{\tau_e(e_1)} v_2 \xrightarrow{\tau_e(e_2)} v_3 \dots \xrightarrow{\tau_e(e_n)} v_n$ , 其中 $\tau_e(e_i)$ 表示连接两点的边的类型。显然,元路径是由一组不同类型的边与所连接的节点构成的路径,这与多元关系事实的表达相一致。另外,图卷积神经网络采用端到端的模式学习更新节点表示,其层间前向传播计算公式如下,

$$H^{(l+1)} = \sigma \left( \tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}} H^{(l)} W^{(l)} \right) \quad (3-3)$$

式中 $\tilde{A} = A + I$ 表示包含节点自环的图邻接矩阵, $\tilde{D}$ 表示 $\tilde{A}$ 的度矩阵,即 $\tilde{D}_{ii} = \sum_j \tilde{A}_{ij}$ ,  $\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$ 表示图的静态结构特征,模型实际可学习部分为 $H^{(l)} W^{(l)}$ 。因此,图卷积运算主要由给定的图静态结构决定,同时也限制了图卷积神经网络在一些存在图结构动态变化的场景的应用。

不同于图卷积神经网络,图变换网络能够适应图结构的动态变化,即其在生成元路径的过程中,允许不同网络层上的图结构存在相应变化,因此,其对应公式(3-3)中的 $\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D}^{-\frac{1}{2}}$ 在各网络层上可以不同。下面通过实例:(诺贝尔化学奖,授予,本杰明·利斯特),(授予\_人物,戴维·麦克米伦),(授予\_时间,2021)表示的多元关系事实“2021年诺贝尔化学奖授予本杰明·利斯特和戴维·麦克米伦。”结合多元关系基础异构图表示形式( $\mathcal{G}_{h-base}$ )来说明,如图3-6所示,该多元关系事实根据其中关系实体之间的连接可构成一个图邻接矩阵(下图左侧所示),同时基于

$G_{h-base}$ 表示形式中存在的四种连接类型，可分解得到对应四个子图邻接矩阵（图右侧所示），则根据图变换网络中元路径生成过程，有图 3-7 中的过程。

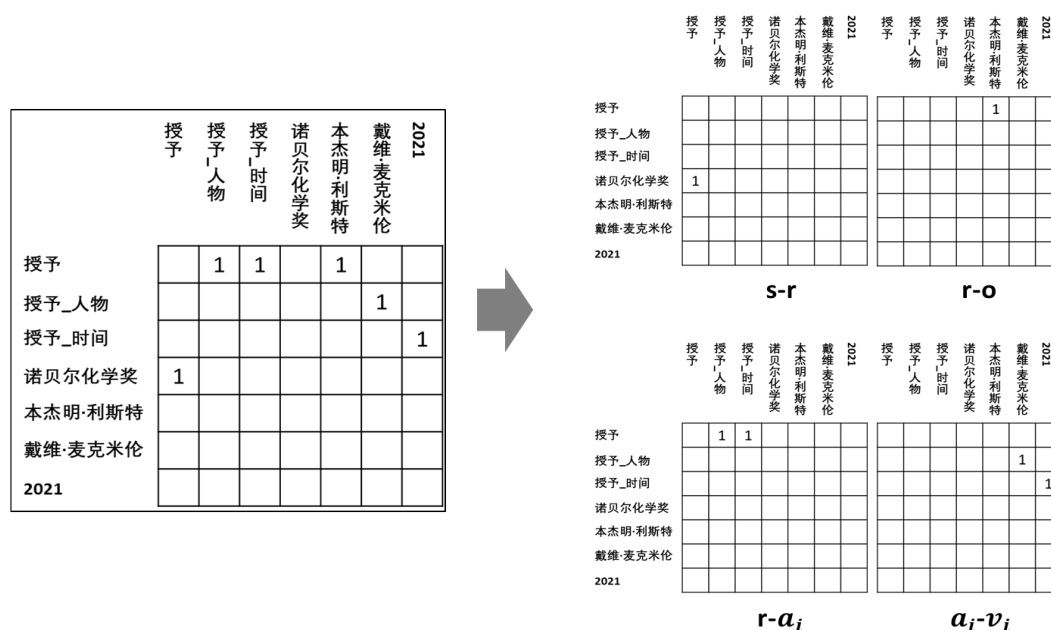


图 3-6 多元关系事实基于实体关系的连接类型分解示意图

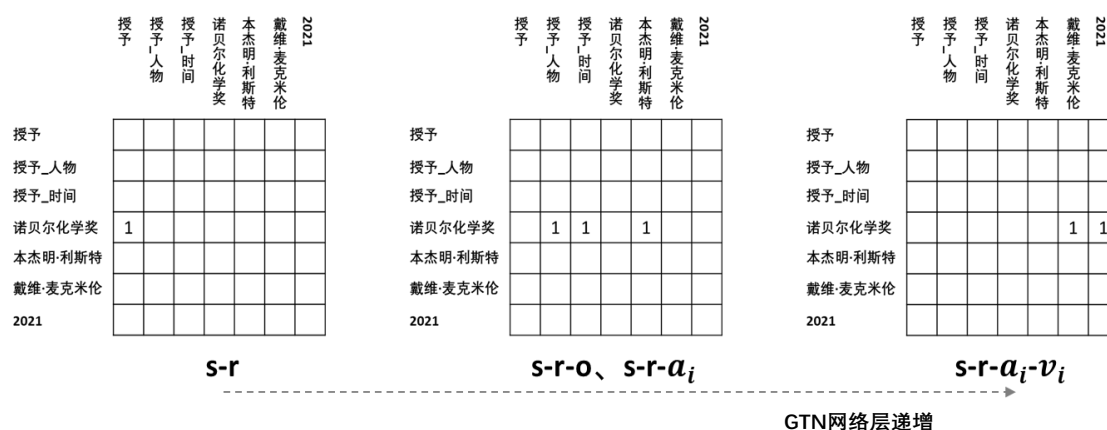


图 3-7 元路径生成过程的简单示意图

从图 3-7 中可观察到，随图变换网络层增加，深层网络上生成的元路径长度递增，图中元路径长度由左至右依次为 1-2-3，且不同网络层上对应元路径图（图中

特征矩阵)存在明显的差异,即图变换网络在各层上的 $\tilde{A}$ 是不同的。因此,其各层上 $\tilde{D}^{-\frac{1}{2}}\tilde{A}\tilde{D}^{-\frac{1}{2}}$ 也不同,即各层网络上图结构动态变化。

图变换网络模型架构如图 3-8 所示。对一个多元关系事实,其异构图表示按图中边的类型分解得到的一组邻接矩阵 $\mathbf{A}$ 作为模型输入。不同于过去需要通过手动定义元路径生成元路径图,然后在元路径图上利用图卷积神经网络学习节点表示,图变换神经网络通过采用网络层上矩阵乘法,实现卷积运算并能够自动查找有效元路径生成元路径图。在图中,模型首先通过自适应筛选出的两个邻接矩阵相乘得到第一层网络上元路径图,之后每次将前一层输出元路径图与自适应筛选出的新输入邻接矩阵相乘来对元路径图进行更新,其效果对应于使得图中所包含的有效元路径长度+1,更新后的元路径图即作为当前层上元路径图。最后通过网络层迭代,从而让元路径图能够表达出多元关系事实中的全部关联信息。

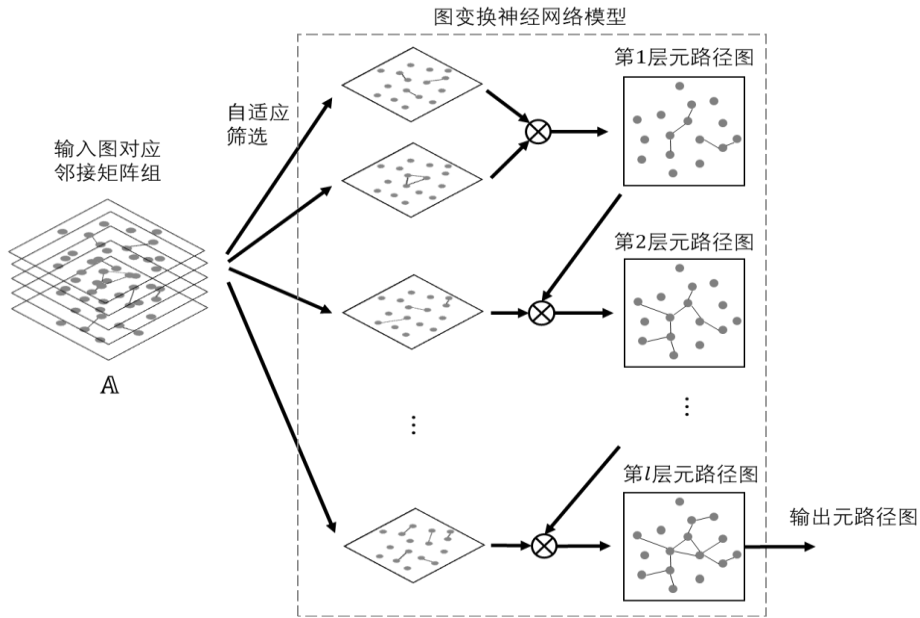


图 3-8 图变换网络模块架构图

模型对 $l$ 长度的元路径所对应元路径图的生成可以表示为,

$$A_{meta-path} = A_{\tau(e_1)} A_{\tau(e_2)} \dots A_{\tau(e_l)} \quad (3-4)$$

式中 $A_{\tau(e_i)}$ 表示路径中第 $i$ 条边所属类型对应的子邻接矩阵。实际训练过程中采用了 $1 \times 1$ 卷积和 Softmax 函数,实现对输入子邻接矩阵组中目标子邻接矩阵的自适应选择,见公式(3-5),参数 $\alpha$ 表示子邻接矩阵之间关联性,长度与 $\mathbf{A}$ 中邻接矩阵数量一致。

$$\mathcal{F}(\mathbf{A}, \phi(\alpha)) = conv_{1 \times 1}(\mathbf{A}, Softmax(\alpha)) \quad (3-5)$$

式(3-5)又可表示为,

$$\mathcal{F}(\mathbb{A}, \phi(\alpha)) = \sum_{t=\tau(e_1)}^{t=\tau(e_l)} \text{softmax}(\alpha_t) A_t \quad (3-6)$$

式中 $\alpha_t$ 对应 $1 \times 1$ 卷积核中具体关联权值。因此, 可得元路径图的更新公式,

$$A^{(l)} = (\tilde{D}^{(l)})^{-1} A^{(l-1)} \mathcal{F}^{(l)}(\mathbb{A}, \phi(\alpha)) \quad (3-7)$$

式中 $A^{(l-1)} \mathcal{F}^{(l)}(\mathbb{A}, \phi(\alpha))$ 表示由前一层上元路径图与当前层上新输入邻接矩阵相乘得到当前层的元路径图(表现为特征矩阵), 网络初始层为 $A^{(0)} = \mathcal{F}^{(0)}(\mathbb{A}, \phi(\alpha))$ 。为保证网络层上数据的稳定性, 式中对当前层上元路径图通过对应的度矩阵 $\tilde{D}^{(l)}$ 进行了行规范化。此外, 为了使模型能够表达任意长度的元路径, 类似于卷积神经网络, 其在 $\mathbb{A}$ 中增加了一个单位阵 $A_0 = I$ 。

借鉴图卷积神经网络的前向传播公式(3-3), 图变换神经网络的表达式为,

$$Z = \sigma(\tilde{D}^{-1}(A^{(L)} + I)XW) \quad (3-8)$$

式中 $A^{(L)}$ 表示图变换神经网络在第 $L$ 层上输出的元路径图,  $\tilde{D}^{-1}$ 为 $(A^{(L)} + I)$ 矩阵的度矩阵, 用于数据规范化,  $X$ 表示输入关系事实包含的所有节点对应特征向量,  $W$ 表示可学习参数矩阵。代入公式(3-7), 可对图变换神经网络表示式展开,

$$\begin{aligned} Z &= \sigma(\tilde{D}^{-1}XW + \tilde{D}^{-1}A^{(L)}XW) \\ &= \sigma(\tilde{D}^{-1}XW + \tilde{D}^{-1}((\tilde{D}^{(L)})^{-1}A^{(L-1)}\mathcal{F}^{(L)}(\mathbb{A}, \phi(\alpha)))XW) \\ &= \sigma(\tilde{D}^{-1}XW + \tilde{D}^{-1}(\tilde{D}^{(L)})^{-1} \dots (\tilde{D}^{(1)})^{-1}A^{(0)}\mathcal{F}^{(1)}(\mathbb{A}, \phi(\alpha)) \dots \mathcal{F}^{(L)}(\mathbb{A}, \phi(\alpha))XW) \end{aligned} \quad (3-9)$$

为了简化计算, 文献[21]中给出并证明了三个等式, 即对给定的两个规范化邻接方阵 $A, B \in R^{N \times N}$ , 有下面等式成立:

$$(D_A^{-1}A)(D_B^{-1}B) = D_{AB}^{-1}AB \quad (3-10)$$

$$D_{AB}^{-1} = I \quad (3-11)$$

$$D_{A+I}^{-1} = (D_A + I)^{-1} = \frac{1}{2}I \quad (3-12)$$

式中由于邻接方阵 $A, B$ 都是行规范化, 各行之和为 1, 易知 $AB$ 也是行规范化矩阵, 从而易推出 $D_A^{-1}, D_B^{-1}, D_{AB}^{-1}$ 都是单位阵。因此, 只需式(3-7)中的 $\mathcal{F}^{(l)}(\mathbb{A}, \phi(\alpha)), l = 0 \dots L$ 事先行规范化, 即可将式(3-9)简化为,

$$Z = \sigma\left(\frac{1}{2}XW + \frac{1}{2}A^{(0)}\mathcal{F}^{(1)}(\mathbb{A}, \phi(\alpha)) \dots \mathcal{F}^{(L)}(\mathbb{A}, \phi(\alpha))XW\right) \quad (3-13)$$

为增加鲁棒性, 文献[21]用超参数 $\gamma$ 替换上式中 $1/2$ , 因此, 得到图变换网络的计算公式(3-14),

$$Z = \sigma((1 - \gamma)XW + \gamma A^{(0)}\mathcal{F}^{(1)}(\mathbb{A}, \phi(\alpha)) \cdots \mathcal{F}^{(L)}(\mathbb{A}, \phi(\alpha))XW) \quad (3-14)$$

通过上面公式可减少模型计算量, 更快训练模型收敛, 得到融合图语义结构特征的节点特征嵌入。

### 3.3.4 目标节点预测过程

目标节点预测过程通过堆叠两层全连接神经网络 FCN 及计算相似度实现。对任一待测关系事实实例, 通过索引从之前输出的  $Z$  中提取其所包含关系实体的对应特征嵌入, 对齐拼接后得到特征矩阵  $Z_k$  作为 FCN 输入, 具体计算公式如下,

$$Pred = Linear_2(\sigma Linear_1(Z_k)) \quad (3-15)$$

$$P = Similar(Pred, Z_{truth}) \quad (3-16)$$

式(3-15)中,  $\sigma$  表示双曲正切激活函数  $\tanh$ , 式(3-16)中预测估计分值  $P$  通过计算目标预测值与真实值  $Z_{truth}$  的相似度得到, 这里相似度采用向量乘法实现, 在所有节点嵌入的数值分布相对均匀的条件下, 节点预测值与真实值越接近, 其乘积的数值越大。

模型采用交叉熵损失函数, 计算公式为,

$$Loss = -\frac{1}{N} \sum_i \sum_{c=1}^C y_{ic} \log(p_{ic}) \quad (3-17)$$

式中  $N$  表示预测节点数量,  $C$  表示类别数量, 模型将每个不同节点视为一个类别,  $y_{ic}$  为 0-1 示性函数,  $p_{ic}$  为预测节点的评估分值。

## 3.4 实验设计与结果分析

本节中, 使用在第二章中介绍的 JF17K 数据集对本章提出的方法展开对比实验。首先介绍实验环境、采用数据集和评估指标, 然后具体介绍实验设计以及相应模型参数设置, 最后将实验结果进行对比并展开分析, 从而验证本章所提出方法的有效性。本节中共设计两组对比实验, 分别对比了本章提出的多元关系异构图表示对比其他两种变体的合理性, 以及提出的基于多元关系异构图表示的图变换网络预测方法相比现有方法在一定程度上的优越性。

### 3.4.1 实验环境与数据

实验环境所使用的详细配置如表 3-1 所示。

表 3-1 实验环境配置表

环境	配置
CPU	Intel(R) Core(TM) i5-10400F CPU @ 2.90GHz
CPU 内存	16GB
GPU	GeForce GTX 1080Ti
GPU 显存	11G
OS	Ubuntu
Cuda 版本	10.2
Pytorch 版本	1.9.0
Python 版本	3.7.10

实验采用的数据集 JF17K 相关参数如表 3-2 所示（本文中数据集版本来自文献[2]）。为更好的完成实验，本文基于对多元关系的理解以及数据集 JF17K 中关系事实的特点为每个关系事实样本添加了预测标签。

表 3-2 数据集 JF17K 参数

数据集	JF17K
实体数/Entities	28645
关系数/Relations	322
训练集样本数	76379
Train Facts	多元关系事实样本数(N>2): 32169 (42.2%)
测试集样本数	24568
Test Facts	多元关系事实样本数(N>2): 14151 (57.6%)

### 3.4.2 实验评估指标

在多元关系链路预测问题中，对潜在关系与实体的预测类似于从一组候选者中推荐出一个最佳结果，现有研究所采用的评估指标包括平均倒数排名 MRR 和 Hits@K（K=1,3,5,10）这两类推荐算法评价指标，为更好的与现有方法进行对比，本文采用相同评估指标方案。

1. 平均倒数排名 MRR，该指标反映预测的正确三元组在所有三元组中的预测分值排名的倒数的平均值，指标值越大越好。

$$MRR = \frac{1}{N} \sum_{i=1}^N \frac{1}{rank_i} \quad (3-18)$$

2. Hits@K，该指标反映对所有预测三元组分值进行排名后，正确三元组排名在前 K 个以内（含第 K 个）的概率，指标值越大越好。

$$Hits@K = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(rank_i \leq K) \quad (3-29)$$

其中 $\mathbb{I}(\text{rank}_i \leq K)$ 表示 indicator 函数, 满足条件 $\text{rank}_i \leq K$ 时函数值等于 1, 否则为 0, 相关文献中多采用  $K=1,3,5,10$ 。

### 3.4.3 实验设计与参数设置

为了验证本章所提出多元关系异构图表示的合理性以及基于多元关系异构图表示的图变换网络预测方法的有效性, 本节基于 JF17K 数据集分别设计了对比实验。通过对比前面提出的三种异构图表示在同一种方法下的预测结果, 验证本文提出的包含实体关系节点、(主/辅) 关系实体对内部连接、主关系与辅助关系连接、辅助关系之间连接的异构图表示的合理性; 通过对比本章提出的预测方法与现有基于一组关系实体对表示的主流多元关系链路预测方法, 验证本章方法在多元关系预测问题中的有效性以及一定程度上的优越性。

在后续实验中, 为表述方便, 这里将本章模型缩写为 NHGTN (N-ary relational Heterogeneous Graph Transformer Network), 其参数如下表进行设置,

表 3-3 模型相关参数设置

参数	取值
Batch-size	256
Grade accumulation steps	6
LR (Model-1)	0.001/relation, 0.05/entity
LR (Model-2)	0.001
LADIES sampling depths	5
NHGTN layers	7

为了避免模型在训练初期出现震荡, 提高模型的稳定性和收敛速度, 训练中学习率更新采用了 Warm-Up 方法, 并设置学习率在初始的 6 次迭代线性增长达到预设值, 随后呈指数衰减。

### 3.4.4 实验结果与分析

#### 1) 多元关系异构图表示的合理性

多元关系异构图表示的合理性验证实验主要由通过三种表示形式 ( $\mathcal{G}_{h-base}$ ,  $\mathcal{G}_{h-std}$ ,  $\mathcal{G}_{h-full}$ ) 在图变换网络链路预测模型上分别进行的实体预测与关系预测两个子实验构成。通过对比 $\mathcal{G}_{h-std}$ 与 $\mathcal{G}_{h-base}$ 的预测结果, 探讨多元关系中辅助关系实体对之间的关联对预测结果产生的影响; 通过对比 $\mathcal{G}_{h-std}$ 与 $\mathcal{G}_{h-full}$ 的预测结果, 探讨多元关系中主体三元组的结构重要性与唯一性。实验结果如表 3-4、表 3-5 所示。

表 3-4 基于三种异构图表示的 NHGTN 模型的关系预测结果

模型	关系 MRR	关系 Hits@1	关系 Hits@3	关系 Hits@5	关系 Hits@10
$\mathcal{G}_{h-base}$	<b>0.9612</b>	0.9385	<b>0.9838</b>	<b>0.9886</b>	<b>0.9917</b>
$\mathcal{G}_{h-std}$	0.9610	<b>0.9399</b>	0.9813	0.9876	0.9914
$\mathcal{G}_{h-full}$	0.9397	0.9097	0.9683	0.9807	0.9869
模型	主关系 MRR	主关系 Hits@1	主关系 Hits@3	主关系 Hits@5	主关系 Hits@10
$\mathcal{G}_{h-base}$	0.9590	0.9348	<b>0.9829</b>	<b>0.9881</b>	<b>0.9914</b>
$\mathcal{G}_{h-std}$	<b>0.9613</b>	<b>0.9408</b>	0.9814	0.9870	0.9911
$\mathcal{G}_{h-full}$	0.9500	0.9259	0.9732	0.9810	0.9870

表 3-5 基于三种异构图表示的 NHGTN 模型的实体预测结果

模型	实体 MRR	实体 Hits@1	实体 Hits@3	实体 Hits@5	实体 Hits@10
$\mathcal{G}_{h-base}$	0.4465	0.2701	0.5423	<b>0.6965</b>	<b>0.8210</b>
$\mathcal{G}_{h-std}$	<b>0.4551</b>	<b>0.2936</b>	<b>0.5448</b>	0.6665	0.7890
$\mathcal{G}_{h-full}$	0.4308	0.2663	0.5279	0.6554	0.7735
模型	主实体 MRR	主实体 Hits@1	主实体 Hits@3	主实体 Hits@5	主实体 Hits@10
$\mathcal{G}_{h-base}$	0.4525	0.2738	0.5502	<b>0.7065</b>	<b>0.8342</b>
$\mathcal{G}_{h-std}$	<b>0.4642</b>	<b>0.3017</b>	<b>0.5555</b>	0.6770	0.8014
$\mathcal{G}_{h-full}$	0.4364	0.2715	0.5348	0.6638	0.7791

表 3-4 与 3-5 中的主关系、主实体分别是指主体三元组中关系、头尾实体，而关系、实体则分别指关系事实中包括主体三元组的实体关系在内的混合关系、混合实体。需要说明，由于现有研究中主要讨论的是在主关系实体的预测能力或者是在混合关系实体的预测能力或者兼而有之，因此本文方法与现有方法的预测结果讨论也从主关系实体、混合关系实体展开，具体而言，对非基于主体三元组与辅助关系实体对表示形式建立的方法主要对比混合关系实体的预测结果，反之则主要对比主关系实体的预测结果，使得得到的分析结论更具有代表性。

观察整体实验结果，在实体与关系的预测上存在明显差异，其原因在于数据集 JF17K 中实体关系数量的巨大差距，其中近 3 万个实体与 300 多个关系，比例接近 100 : 1，但这不会影响在预测模型的分析结果。

关系事实中主体三元组具有语义主体性与结构唯一性。观察  $\mathcal{G}_{h-std}$  与  $\mathcal{G}_{h-full}$  两种表示形式的实验结果，可发现前者在实体关系预测准确性上全面优于后者，其中指标 Hits@1 在主关系预测提升了 1.60%、在主实体预测提升了 11.12%；指标 MRR 在主关系预测提升了 1.19%、在主实体预测提升了 6.37%。因此，可推断  $\mathcal{G}_{h-full}$  中



增加的主体三元组中头尾实体与辅助关系实体对的连接实际降低了对目标节点的预测准确性。其原因在于，这两类连接实际是从语义结构上重构了分别由头尾实体到辅助关系实体对的三元组结构，使得在一个关系事实中出现了多个三元组嵌套，其破坏了关系事实中原本主体三元组的结构唯一性特征，语义上模糊了主体三元组高于辅助关系实体对的重要性优先级。 $\mathcal{G}_{h-full}$ 表示形式从结构上实际是一组关联的平级三元组表示结构，其在效果上也实际增加了模型的训练难度。

关系事实中当辅助关系实体对之间存在语义依赖时，基于 $\mathcal{G}_{h-std}$ 的模型能够得到更高的预测准确率。观察 $\mathcal{G}_{h-std}$ 与 $\mathcal{G}_{h-base}$ 两种表示形式的实验结果，可发现从整体上两者各有优劣，而从预测的准确性角度， $\mathcal{G}_{h-std}$ 显然相对更有优势，其中在指标 Hits@1，主关系预测提升了 0.64%、主实体预测提升了 10.19%；指标 MRR 在主关系预测提升了 0.24%、在主实体预测提升了 2.59%。因此，可推断 $\mathcal{G}_{h-std}$ 相对于 $\mathcal{G}_{h-base}$ ，前者增加的辅助关系实体对之间的连接能够帮助提升目标节点的预测准确性，其原因在于当关系事实的辅助关系实体对之间存在语义依赖时， $\mathcal{G}_{h-std}$ 帮助模型提供更多的特征信息，从而能够提高对目标的预测准确性。但同时也观察到 $\mathcal{G}_{h-base}$ 却能在指标 Hits@5，Hits@10 有稳定的优势，其原因在于关系事实本身，即并非所有关系事实的辅助关系实体对之间存在语义依赖，当辅助关系实体对之间在语义结构上相对独立， $\mathcal{G}_{h-std}$ 增加的连接实际转变为一种噪声，从而降低了预测结果。然而， $\mathcal{G}_{h-base}$ 在指标 Hits@10 的主关系预测提升 0.03%、主实体预测提升 4.09%均小于 $\mathcal{G}_{h-std}$ 在指标 Hits@1 的对应提升幅度，因此，从整体上仍可认为 $\mathcal{G}_{h-std}$ 相比 $\mathcal{G}_{h-base}$ 具有更符合多元关系事实的表示形式优势。

上述实验结果，证明了本章所提出的多元关系的异构图表示形式在建模多元关系事实的可行性与有效性，以及所提出的 NHGTN 模型在多元关系链路预测任务中的可行性。

## 2) NHGTN 模型案例分析

为了进一步分析 NHGTN 模型的预测过程，以多元关系事实：“2021 年诺贝尔化学奖授予德国科学家本杰明·利斯特（Benjamin List）和美国科学家戴维·麦克米伦（David MacMillan）”为例，将其中“诺贝尔化学奖”和“授奖”分别作为实体、关系预测对象，模型对于目标实体的预测过程以及在不同网络层数的预测正确结果所在（Top10）排名如图 3-9 所示，对应实体、关系的预测结果（Top5）如表 3-6 所示。

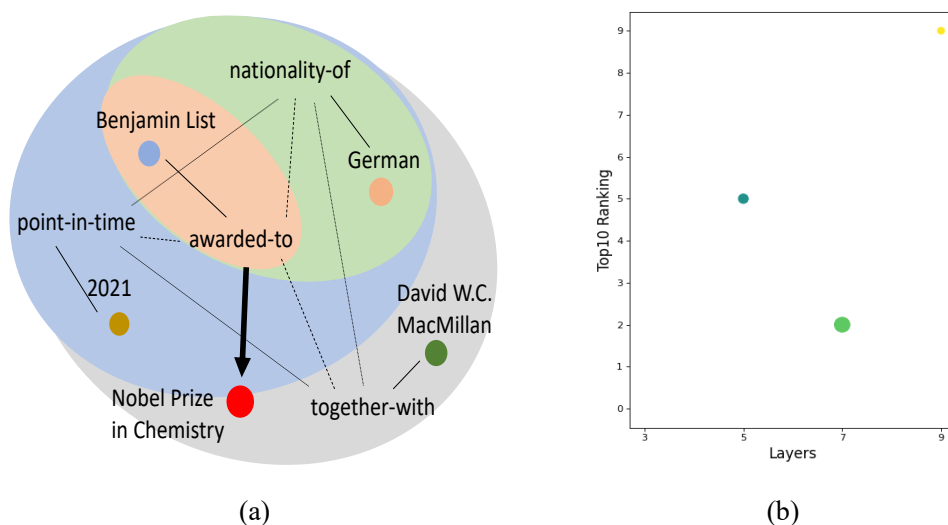


图 3-9 NHGTN 模型的预测过程。(a)案例中对主实体“诺贝尔化学奖”的预测过程；(b)模型在不同层时，预测出正确结果的 Top10 排名

表 3-6 NHGTN 模型（7 Layers）对实体“诺贝尔化学奖”、关系“授予”的预测结果

实体预测	<i>Nobel Prize in Chemistry</i> is awarded to German scientist Benjamin List and Scotland-born scientist David W.C. MacMillan of Princeton University in 2021.				
	1. Wolf Prize in Chemistry	2. <i>Nobel Prize in Chemistry</i>	3. Nobel Prize in Physics	4. Priestley Medal	5. The Abel Prize
关系预测	The 2021 Nobel Prize in Chemistry is <i>awarded to</i> German scientist Benjamin List and Scotland-born scientist David W.C. MacMillan of Princeton University.				
	1. <i>awarded-to</i>	2. award-nominated	3. announced-to	4. made-by	5. honored-for

图 3-9(a)表示对主实体“诺贝尔化学奖”的预测过程，其中体现了事实异构信息的聚合：圆表示的主实体“本杰明·利斯特”经实线连接椭圆表示的主关系“授予”、主关系与辅助关系“国籍”虚线连接，和辅助实体“德国”串联，对应三层网络模型的输出元路径图；再由“授予”与辅助关系“时间点”连接、以及辅助关系“国籍”与“时间点”之间点线连接，关联“2021”，对应五层网络模型的输出元路径图；最后同理关联“戴维·麦克米伦”，对应七层网络模型的输出元路径图。图 3-9(b)反映了对应层数不同的模型预测出正确实体的所在排名。

在预测过程中，实体关系节点异构特征聚合，表现为椭圆区域面积逐层增大，因此，模型最后的正确预测结果可以认为出现在整体区域面积之中。在图(a)的特征聚合过程中，看到正确结果有时在较小面积区域中出现，对应较少网络层数的模型也存在正确预测出结果的可能，如图(b)中模型在五层时预测出正确结果并且排

在第 5 位,然而模型在七层网络时,因聚合了完整事实信息,其预测出正确结果的排名更高排在第 2 位。因此,通过案例,完成了对 NHGTN 模型预测过程的分析。

### 3) NHGTN 预测模型的有效性

为了验证 NHGTN 模型对比当前基于一组关系实体对表示的典型预测方法具有一定的竞争力,将基于  $\mathcal{G}_{h-std}$  的 NHGTN 模型实验结果与现有方法在混合关系、实体预测上分别进行对比,其结果如表 3-7、表 3-8 所示。

NHGTN 模型在多元关系的混合关系预测上具有明显优势。观察表 3-7,对比次优模型 tNaLP<sup>+</sup>, NHGTN 在所有指标上都有不同程度的提升,说明 NHGTN 在关系预测方面具有明显的优势,其多元关系的异构图表示能够充分表现多元关系事实中的隐藏信息,并通过图变换网络提取和融合,实现对事实中关系的准确预测。而在具体指标上, NHGTN 分别在 Hits@1 提升了 0.20%、在 Hits@3 提升了 1.06%、在 Hits@10 提升了 0.55%,其中 Hits@3 和 Hits@10 的提升明显更高,说明模型对于混合关系预测准确性的提升,主要是从增加预测相对准确的关系样本数量的角度,即 NHGTN 模型能够对更多数量的待测样本提供相对准确的预测,从直觉上,这也贴合实际的推荐应用,比如在搜索引擎中,对于任意用户查询,更大可能是在一个可接受范围内输出用户满意的结果,通常前几个输出结果对于用户的需求高度接近,难以直接断定第一个结果就能是用户所需。因此,可以认为 NHGTN 模型在指标 Hits@3、Hits@10 的提升相比 Hits@1 具备同样甚至更多重要性。

表 3-7 NHGTN 模型与基于一组关系实体对表示的现有模型的关系预测对比结果

模型	关系 MRR	关系 Hits@1	关系 Hits@3	关系 Hits@10
NaLP (2019)	0.825	0.762	0.873	0.927
tNaLP <sup>+</sup> (2021)	0.956	0.938	0.971	0.986
NHGTN( $\mathcal{G}_{h-std}$ )	<b>0.9610</b>	<b>0.9399</b>	<b>0.9813</b>	<b>0.9914</b>

表 3-8 NHGTN 模型与基于一组关系实体对表示的现有模型的实体预测对比结果

模型	实体 MRR	实体 Hits@1	实体 Hits@3	实体 Hits@10
NaLP (2019)	0.366	0.290	0.391	0.516
tNaLP (2021)	0.384	0.315	0.411	0.513
tNaLP <sup>+</sup> (2021)	0.449	0.370	0.484	0.598
RAM (2021)	<b>0.539</b>	<b>0.463</b>	<b>0.573</b>	0.690
NHGTN( $\mathcal{G}_{h-std}$ )	0.4551	0.2936	0.5448	<b>0.7890</b>

在多元关系的混合实体预测, NHGTN 模型在一定程度上仍然具有竞争力。从表 3-8 中,对比 RAM 模型, NHGTN 在指标 Hits@10 有 14.35%的明显提升,说明

模型确实能够为更多数量的样本提供相对准确的预测结果，这对于允许输出多个预测结果的推荐类应用更有吸引力，因正确结果出现的可能性相对更大。此外，相比同时期的 tNaLP<sup>+</sup>模型，NHGTN 还在 Hits@3 和 MRR 上分别提升了 12.56%、1.36%。因此，可认为 NHGTN 模型在实体预测中，一定程度上仍具有竞争力。

上述实验结果，证明了图变换网络的思路应用在大规模图谱的多元关系数据集上的可行性和有效性，并对比现有模型具备一定程度上的竞争力；同时，也证明了保留主体三元组结构特征的多元关系异构图表示形式与关系事实的语义结构特征相符。而在实验中发现 NHGTN 模型在指标 Hits@1 上的预测结果，尤其在实体预测上相对偏低，其原因可能来自于图变换网络本身在元路径图生成过程中存在的信息丢失问题，这将在下一章给出详细分析和实验证明。

### 3.5 本章小结

本章针对多元关系链路预测问题提出了一种基于多元关系异构图表示的图变换网络预测方法。首先，基于对多元关系事实的理解，分析了现有多元关系的不同表示形式：传统二元关系预测方法中对多元关系的处理、多元关系的基于一组关系实体对表示以及基于主体三元组与辅助关系实体对表示，并说明了各表示形式所存在的一些问题。然后针对基于一组关系实体对表示形式的现有一些链路预测方法的特点和存在的问题进行了重点分析，并根据主体三元组与辅助关系实体对表示和对多元关系的理解提出了面向多元关系的异构图表示。为了充分挖掘和利用异构图上关系事实信息，本章提出引入图变换网络模块，利用其元路径生成过程实现图异构信息提取和聚合。为了让图变换网络能够适用于多元关系链路预测问题中常用的大规模图数据集，本章进一步引入并适应性调整了 LADIES 采样模型，同时，在实验过程中通过为关系事实增加一个降噪模块，增强了多元关系异构图表示在不同关系事实上的表达能力。综合上述内容提出了本章基于多元关系异构图表示的图变换网络预测模型 NHGTN，并介绍了其模型架构和运行机制等。最后在公开数据集 JF17K 进行了对比分析实验，结果证明了本章提出的多元关系异构图表示具有充分表达多元关系事实信息的能力，也证明了提出的 NHGTN 模型算法思路在多元关系预测问题上的可行性和有效性。

## 第四章 基于轻量化聚合的增强图变换链路预测方法研究

受不同长度变化的多元关系事实影响，图变换网络在生成元路径图的过程中，深层网络存在信息丢失的问题。为此，在前一章模型的基础上，本章提出了一种轻量化聚合方法，将图变换网络中所有网络层视为通道，构建了跨通道图最大池化的特征聚合方案，实现了对模型输出层上对分散各层的关键特征聚合。本章首先对图变换网络在深层网络的信息丢失问题进行了分析，并分析介绍了基于主体三元组与辅助关系实体对表示的主流链路预测方法，然后结合对图变换网络各层上元路径特征图的分析，提出了基于轻量化聚合的增强图变换链路预测方法的建模思路和模型架构，最后在公开数据集 JF17K 上设计了对比分析实验，结果表明本章方法在 MRR 指标上，对比 HINGE 模型在主关系预测提升了 1.37%，主实体预测提升了 13.88%。

### 4.1 引言

在一个多元关系数据集中，不同关系事实所包含的关系、实体的数量通常不是完全相同，即关系事实的长度不同。因此，在对数据集中所有关系事实样本按照不同的连接类型建立图邻接矩阵时，部分关系事实可能在其中对应某些连接类型的邻接矩阵上没有值，或者说相比其他关系事实所包含的全部连接类型的邻接矩阵，这些事实只含有其中部分邻接矩阵。

图变换网络模型是利用前一层输出的元路径图与筛选出的邻接矩阵通过矩阵乘法生成当前层上元路径图，使得图中的包含的元路径长度递增，因此，当模型随着网络逐层加深，其中较短的关系事实由于长度限制，在深层网络上不可避免将被自动“遗忘”，即存在信息丢失。

下面以一个例子来简要说明该问题，给定一组多元关系事实，其中关系、实体统一用节点  $U = \{a, b, c, d, e, f\}$  来表示，这组关系事实有，

1.  $a \rightarrow c \rightarrow d \rightarrow b \rightarrow f$
2.  $c \rightarrow f \rightarrow b$
3.  $d \rightarrow e \rightarrow a$
4.  $b \rightarrow c \rightarrow e \rightarrow d \rightarrow f$
5.  $f \rightarrow a \rightarrow d \rightarrow c$

令其中同一关系事实中任意两节点之间连接 ( $\rightarrow$ ) 类型不同、不同关系事实在同一位置上节点间连接类型相同，则关系事实中共有 4 种关联类型，对应 4 个不同类型邻接矩阵。为了简化起见，这里只区分节点之间的连接有无，即在邻接矩阵

中节点之间存在连接则记为 1，没有则为 0（省略）。则根据图变换神经网络的基本原理，有下面计算过程，

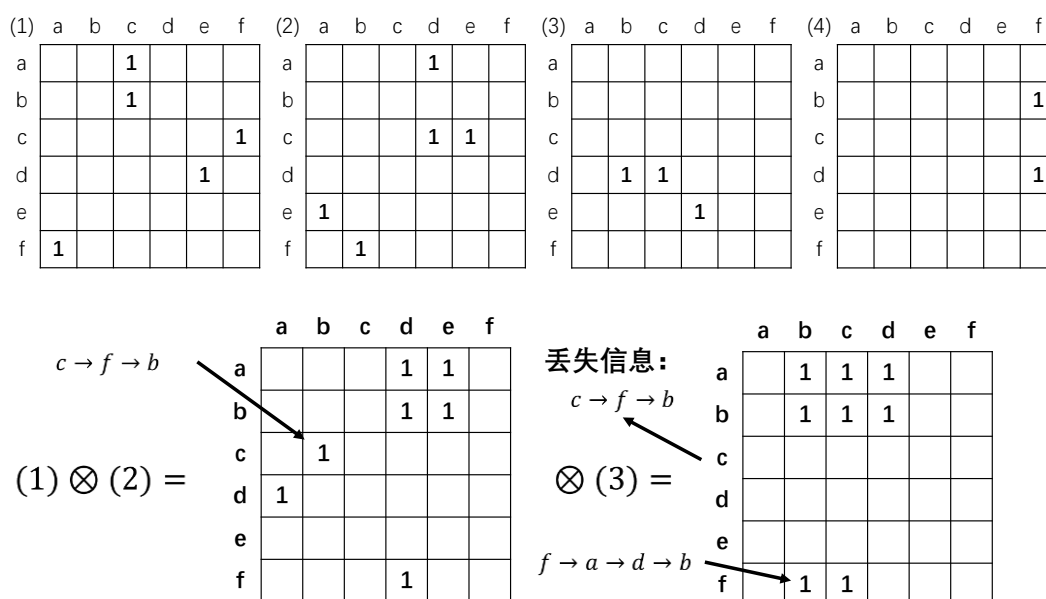


图 4-1 图变换网络上元路径图生成过程

在图 4-1 中可以看出，随着元路径图中元路径长度的递增，之前存在的较短长度的多元关系事实信息，如  $c \rightarrow f \rightarrow b$  将被自动丢失，从而降低了最终模型预测的准确性，这对于一些关系事实长度差异较大的数据集而言尤为不利。

另外，上一章中 NHGTN 模型在实验中发现其模型复杂度偏高，其空间复杂度与数据集中关系实体数量成正比，这对于大规模数据集而言较为不利，并且模型中图采样的输入与小批量学习方法关联，而样本批次数量又与数据集样本数呈正相关，因此当数据集中样本数量过大时，模型训练时间相对偏长，如上一章  $G_{h-std}$  表示的模型在实体预测实验中，单个迭代的训练平均时间约为 50min，120 次迭代的总训练时长接近 100h，其中在 32% 左右时长达到最终预测结果的 90%。

针对上述图变换网络因元路径长度增长而在深层网络存在的信息丢失问题，本章在 NHGTN 模型基础上，结合该模型复杂度相对较高不宜过多增加训练时间的情况，提出了一种轻量化聚合方法，通过将各层网络通道对齐，然后利用跨通道最大池化提取分布在各通道上的最大特征值信息，并聚合到网络输出层，实现各层上关键特征聚合，补全丢失信息，从而提升整体模型的预测效果。

## 4.2 基于主体三元组与辅助关系实体对及其他表示的相关方法分析

在多元关系的链路预测研究中,除了一组关系实体对的表示形式,主体三元组与辅助关系实体对也是一类常见表示形式,如前章所述,主体三元组结构体现了关系事实中内容主体的重要性,因而其表示形式相对更符合事实的语义结构特征,在预测实验中得到相对有更好的结果。

NeuInfer<sup>[10]</sup>模型是基于主体三元组与辅助关系实体对表示形式的典型模型,其完全由全连接神经网络架构组成,包括两条计算执行路线。一条路线上通过两个全连接神经网络 FCN,分别实现主体三元组内部关系实体之间交互特征提取与聚合操作,然后计算该三元组的正确性评估分值;另一条路线同样通过两个 FCN,实现主体三元组与每个辅助关系实体对之间交互特征提取,并在最小池化聚合各组提取后,实现计算主体三元组与各辅助关系实体对的兼容性评估分值。最后综合正确性与兼容性,输出该关系事实中目标节点的对应预测评估分值。整体模型架构简单直观,但其忽略了辅助关系实体对之间可能存在的信息交互,也没有考虑关系事实本身的异构性问题,同时大量的 FCN 结构实际增加了模型的计算开销。

与 NeuInfer 同时期提出的 HINGE<sup>[2]</sup>模型拥有相似的模型架构,其同样将多元关系事实中主体三元组与辅助关系实体对分成两条路线处理,分别提取主体三元组内实体关系特征交互,以及将主体三元组与每个辅助关系实体对重构为五元组的特征矩阵,然后提取其中三元组与辅助关系实体对之间的交互特征。不同之处在于 HINGE 首先将一组一维节点嵌入拼接成二维矩阵,再通过卷积神经网络 CNN 对特征矩阵实现交互信息提取。模型将两条路线上输出特征向量对齐拼接成一个特征矩阵后,再通过最小池化得到对应表示该多元关系事实的特征嵌入,最后经过一个 FCN 计算目标的预测评估分值。HINGE 主要通过采用 CNN 实现了节点交互特征提取,相比 NeuInfer 完全采用 FCN,在计算效率方面进行了提升,但模型同样忽视了辅助关系实体对之间的交互,也没有考虑到关系事实的异构性特征。

多元关系链路预测问题中,还有一些方法并没有明确提出对于多元关系的表示形式,其中部分<sup>[12]</sup>认为多元关系本身复杂抽象,其内部交互难以进一步分解成更小单元。因此,这些方法通过将多元关系建模成高维嵌入,然后或者通过数理几何空间(如双曲空间中 Poincaré Embedding)、或者通过张量分解方法(如 S2S<sup>[13]</sup>模型)实现对整体多元关系事实的计算,并对其中的目标节点进行预测评估。然而,这些方法大多只实现了对其中实体、甚至部分实体的预测,完全没有顾及关系预测,因此在方法的通用性方面存在一定不足。同时,这些方法通常也没有考虑关系事实的异构性。

针对上述基于主体三元组与辅助关系实体对表示的预测方法以及其他方法所存在的问题,本章提出了基于轻量化聚合的增强图变换链路预测方法,其优点在于: 1) 保留了图变换网络对于多元关系事实中实体关系节点的异构特征挖掘能力,并对实体与关系具有同等预测能力; 2) 采用了跨通道图最大池化的轻量化特征聚合方法,实现了在不额外增加模型参数的基础上有效补全模型输出层关键特征,从而有效避免了图变换深层网络上的信息丢失问题,提升了整体的预测效果。

### 4.3 基于轻量化聚合的增强图变换链路预测模型

针对前面的分析中,图变换网络在深层网络上存在的信息丢失问题,结合前面基于多元关系异构图表示的图变换链路预测方法现有复杂度相对较高的情况,本节遵循图变换网络中元路径图生成过程,提出一种轻量化聚合方法,实现了对各层上关键特征的聚合,从而一定程度上克服了图变换网络的信息丢失问题。

#### 4.3.1 总体模型框架

本节提出了轻量化聚合的增强图变换链路预测方法,模型整体框架如图 4-2 所示,其在前一章 NHGTN 模型的基础上主要增加了基于跨通道图最大池化的轻量化聚合模块。

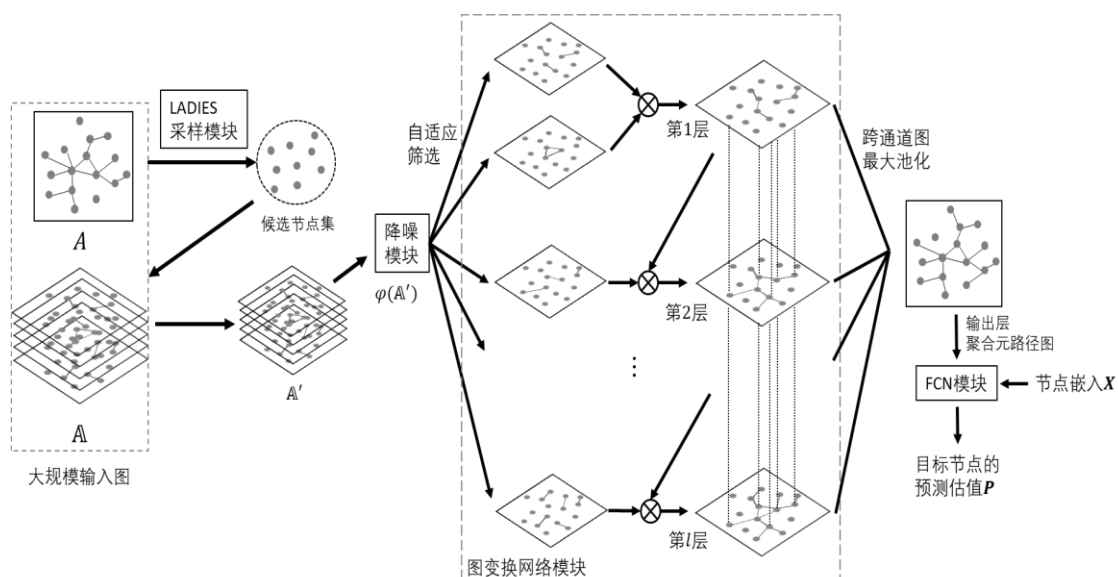


图 4-2 基于轻量化聚合的增强图变换链路预测模型整体架构图

以一组多元关系事实作为输入,模型中主要的图变换网络及轻量化聚合模块的执行过程如下:



(1) 图变换网络的每一层上, 从 $\varphi(\mathbf{A}')$ 自适应筛选出的强关联性子邻接矩阵与上一层输出元路径图通过矩阵乘法生成当前层上的元路径图, 然后将该元路径图备份并按照通道对齐与之前各层上输出元路径图堆叠在一起。

(2) 迭代 $l$ 层后, 对堆叠的 $l$ 层输出元路径图采用跨通道图最大池化, 实现对网络各层上关键特征聚合, 并生成模块输出层聚合元路径图, 用于后续目标节点的预测过程。

### 4.3.2 轻量化聚合模块

图变换网络通过逐层更新元路径图, 对其中元路径自行生成, 实现异构图表示的多元关系事实中语义结构特征的提取和聚合, 同时, 也使得网络各层上输出元路径图存在明显的结构差异, 如图 4-3 所示, 图中第一行表示图变换网络在各层上输出元路径图的图形表示, 反映其从原理上, 随着网络加深, 元路径图所包含的元路径数量、长度都会逐层递增, 图拓扑结构及对应的语义关系也逐层复杂。然而, 在第二行对应的实现过程 (这里做了简化, 只用数值 1 来填充矩阵中数值, 该数值实际表示所在层上的同一位置上, 由所在行列节点作为元路径两端点组成的全部元路径对应特征值之和), 由于图卷积运算本身的局限性, 深层网络的特征矩阵上, 一些原本在浅层特征矩阵中存在的数值在随网络加深过程中丢失了, 从而导致深层网络的特征矩阵中特征值数量有明显减少。

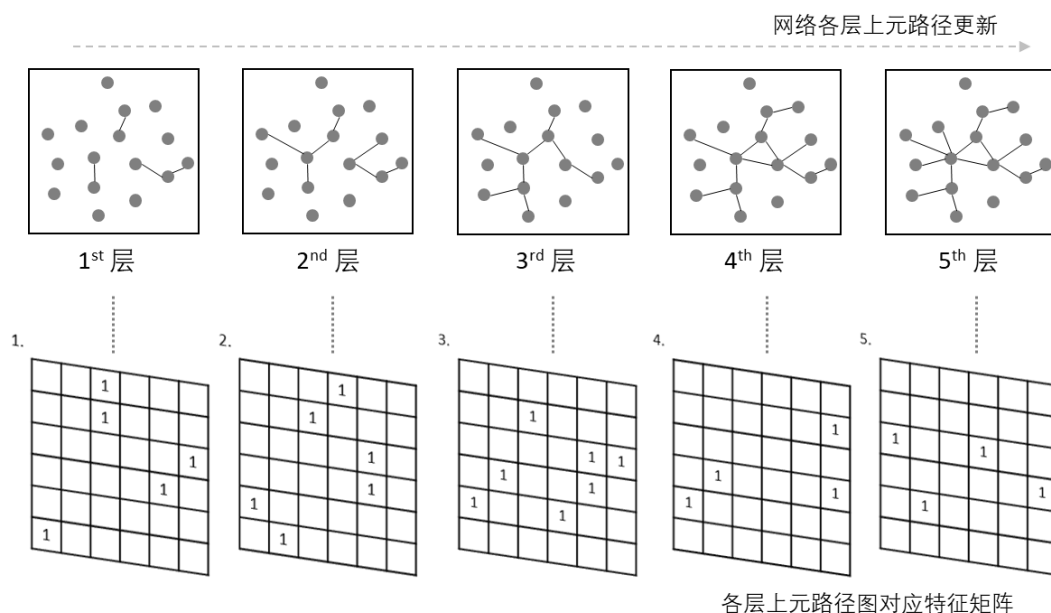


图 4-3 图变换网络各层上元路径图更新过程示意图

为了克服图变换网络的信息丢失问题，本节考虑通过跨网络层聚合的方法将分散在各层上的关键信息聚合，实现输出层特征矩阵补全。由于特征矩阵上数值表示其所在层上由所在行列节点作为元路径两端点组成的所有元路径对应特征值之和，可以认为特征值实际反映了元路径图中任意两个节点之间的关联性强度，即特征值越大，对应两个节点之间的关联性越强；而在同一位置的不同层上，其区别只在于由相同端点组成元路径的长度大小，反映元路径的长度增长。同时，链路预测问题本身是基于潜在目标节点与已知关系事实节点之间必然存在的强关联性，从而利用节点之间的语义结构关联性实现对潜在目标节点的预测，因此在候选节点中与已知节点之间关联性越强的节点应为预测节点。

基于上述分析，可以认为图变换网络中各层输出元路径图对应的特征矩阵，特征值的大小对应节点之间的关联性强度，而在同一位置上，各层中特征值的最大值对应节点之间最大关联性，在链路预测问题中为潜在节点预测提供最准确有效的影响，因此，本节将各层特征矩阵上在同一位置的特征值最大值作为影响潜在节点预测的关键特征。

为了聚合分散在各层网络上的关键特征，本节提出了跨通道图最大池化聚合方案，如图 4-4 所示，在方案中，池化对象为各层输出元路径图对应的特征矩阵，以网络层为通道，经过将特征矩阵按通道对齐，利用图最大池化实现跨通道提取分散的关键特征，并聚合到输出特征矩阵，即输出层聚合元路径图，作为后续潜在目标节点预测过程的输入。

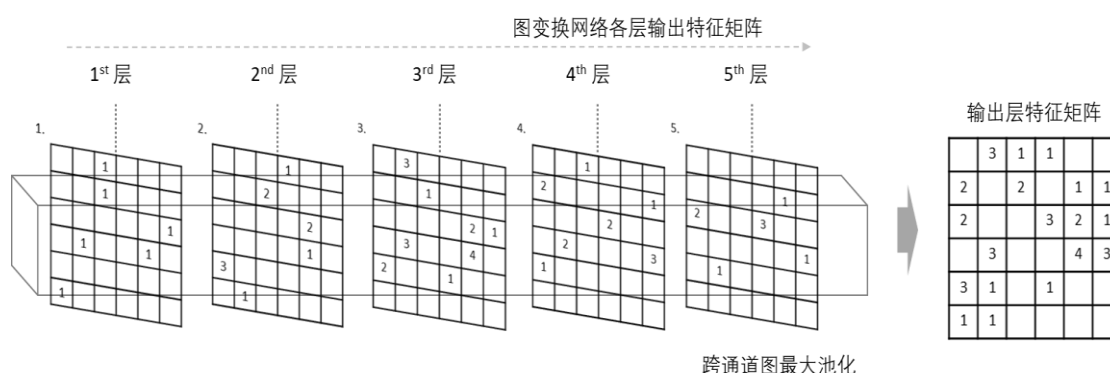


图 4-4 基于跨通道图最大池化的轻量化聚合过程示意图

## 4.4 实验设计与结果分析

本节首先简要介绍了实验环境、相关数据集和评估指标，然后具体介绍了实验设计思路和相关模型参数，最后对实验结果进行了对比分析，从而证明本章方法的有效性。本节共设计三组对比实验，其中在多元关系数据集 WD50K\_100 上对前一章模型在图变换网络存在的信息丢失问题进行验证分析，在第二章中 JF17K 数据集上对本章提出的基于跨通道图最大池化的轻量化聚合方法的合理性与有效性进行了验证分析，以及对本章提出的基于轻量化聚合的增强图变换链路预测模型的有效性与相比现有方法的优越性进行了对比验证。

### 4.4.1 实验环境与数据

本章实验的环境详细配置信息与本文 3.4.1 节完全相同，其详细信息见表 3-1。

实验采用数据集 JF17K 相关信息与上一章相同，下面主要介绍多元关系数据集 WD50K\_100 的相关信息(本文采用数据集版本来自文献[5])，具体信息如下表，

表 4-1 数据集 WD50K\_100 参数

数据集	WD50K_100
实体数/Entities	18792
关系数/Relations	279
训练集样本数	22738
验证集样本数	3279
测试集样本数	5297
多元关系样本总数	31314 (100%)
关系样本的长度分布	Train: 5/7/9/11/13/... = 66.09/28.85/2.87/1.03/0.81/... (%) Valid: 5/7/9/11/13/... = 64.53/30.01/3.26/1.13/0.85/... (%) Test: 5/7/9/11/13/... = 61.96/33.94/2.19/0.68/0.83/... (%)

与上一章相同，为更好的完成实验，本文基于对多元关系的理解以及数据集中关系事实的特点为每个关系事实样本添加了预测标签。

### 4.4.2 实验评估指标

本章实验采用的评估指标与上一章的完全相同。

### 4.4.3 实验设计与参数设置

为了验证 NHGTN 模型关于其图变换网络在深层网络中存在的信息丢失问题，本节设计在完全由多元关系事实样本组成的数据集上对 NHGTN 模型设置不同网络层数，根据层数不同的模型实际输出的预测结果，验证其中图变换网络受多元关

系事实长度变化影响造成在深层网络上存在信息丢失，从而导致模型预测结果偏低的问题。模型相关参数设置如下，

表 4-2 NHGTN 模型关于 GTN 信息丢失问题的相关参数设置

参数	取值
Batch-size	128
Grade accumulation steps	6
LR (Model-1)	0.001/relation 0.05/entity
LR (Model-2)	0.001
LADIES sampling depths	5
NHGTN layers	3/5/7/10

为了验证本章提出的基于跨通道图最大池化的轻量化聚合方法的合理性与有效性、以及进一步验证基于轻量化聚合的增强图变换链路预测模型相比现有方法的有效性与一定程度上优越性，本节后续分别设计了两组对比实验。通过对比基于跨通道图最大池化的轻量化聚合方法与近似平均池化的 $1 \times 1$ 卷积聚合、LSTM 聚合方法的预测结果来验证图最大池化的合理性；通过对比本章所提出预测方法与现有其他方法（包括基于主体三元组与辅助关系实体对表示的部分主流预测方法、基于张量分解等相关方法），验证本章方法在多元关系预测问题中的有效性以及一定程度上的优越性；消融实验中通过与上一章 NHGTN 模型对比，验证轻量化聚合模块的有效性。

在后续实验，为表述方便，这里将本章模型缩写为 LENHGTN (Light Enhanced-NHGTN)，其参数设置如下，

表 4-3 LENHGTN 模型参数设置

参数	取值
Batch-size	256
Grade accumulation steps	6
LR (Model-1)	0.001/relation 0.01/entity
LR (Model-2)	0.001
LADIES sampling depths	5
LENHGTN layers	9/MaxPooling 7/Conv&LSTM

与上一章处理相同，为避免模型在训练初期出现震荡，提高模型的稳定性和收敛速度，两个模型的学习率更新均采用相同的 Warm-Up 方法，使得学习率在初始影响较小且前 6 次迭代线性增长，随后呈指数衰减。

#### 4.4.4 实验结果与分析

##### 1) 验证在图变换深层网络中信息丢失造成的预测结果降低

为了验证图变换深层网络存在信息丢失从而降低了预测结果，本节在 WD50K\_100 数据集上对  $\mathcal{G}_{h-std}$  通过增加图变换的网络层数，从混合实体预测角度对模型在不同网络层数下的预测展开对比实验，对比结果如图 4-5 所示。

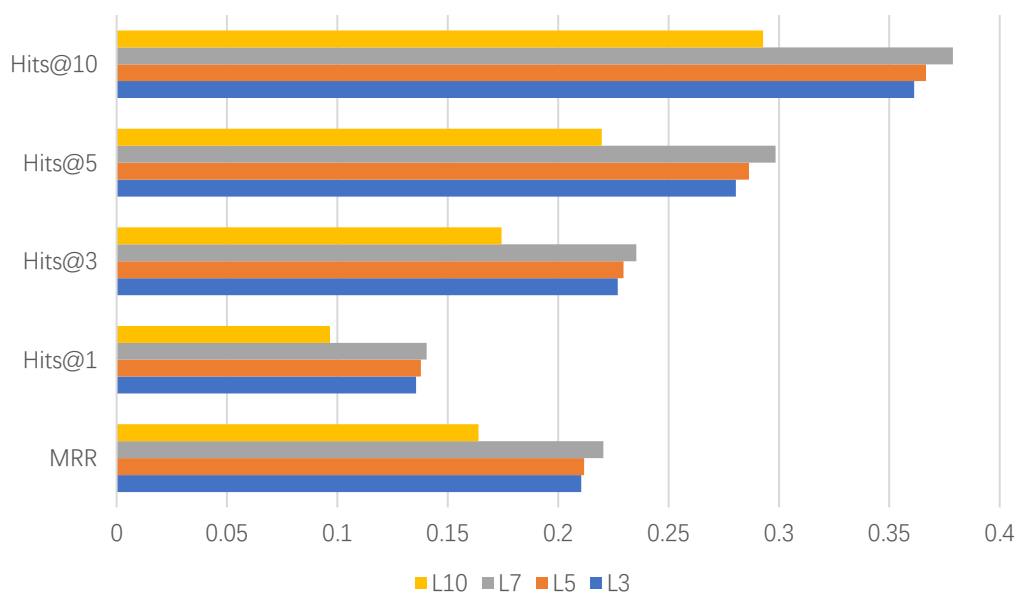


图 4-5 图变换深层网络上信息丢失造成的预测结果降低

可以观察到，随着网络层加深，指标预测结果从整体上呈现出先提升后迅速下降的趋势。模型从 3 层增加到 5 层再到 7 层的过程中预测结果平均分别提升 1.50% 和 3.55%，而在 10 层网络预测结果明显下降，说明数据集的所有关系事实实例对应到整体异构图上，其关系实体节点所构成的有效元路径长度主要为 7，即数据集对应的异构图上任意一条由实体关系节点连接组成的元路径的长度大概率等于 7。

在网络增加到 10 层时，预测结果指标迅速降低，平均为-25.54%。其原因在于网络层数大于异构图上实际有效元路径长度 7，此时在图变换 7 层以上的网络层所对应的元路径图上包含的事实信息反而逐层减少，即存在明显的信息丢失，从而大大降低模型预测的准确性。因此证明了图变换网络确实存在深层网络上信息丢失问题。

有趣的是，从训练数据集上不同长度关系事实实例的分布，可发现长度在 7 以内的关系事实占比超过 98%，其中在 5 以内的占比达到 89%，等于 7 的占比 9.08% 而长度 9 的占比只为 1.83%，结合从实验结果发现异构图上有效元路径长度主要为 7，可认为有效元路径长度与不同长度关系事实的主要占比分布相关，所以图变换

网络层数也应与不同长度关系事实主要占比分布相关。因此,通过上述实验,证明了图变换网络受到不同关系事实长度影响确实存在深层网络上信息丢失问题。

## 2) 验证基于跨通道图最大池化的轻量化聚合方法的合理性与有效性

在 4.3.2 节中分析认为图变换网络的各层输出元路径图对应特征矩阵上,特征值大小对应节点之间的关联性强度,且在同一位置的不同层之间,最大特征值对应节点之间最大关联性,因此将各层特征矩阵上在同一位置的特征值最大值作为影响潜在节点预测的关键特征并提出了基于跨通道图最大池化的轻量化聚合方法。为了验证上述分析的合理性以及所提方法的有效性,本节将其与近似平均池化的  $1 \times 1$  卷积聚合 ( $1 \times 1 Conv$ )、LSTM 聚合 ( $LSTM$ ) 的两种轻量化聚合方法进行对比实验,实验结果如表 4-4、4-5 所示。

表 4-4 三种聚合方法在 LENHGTN 模型的关系预测结果

模型	关系 MRR	关系 Hits@1	关系 Hits@3	关系 Hits@5	关系 Hits@10
$1 \times 1 Conv$	0.9630	0.9436	0.9825	0.9856	0.9883
$LSTM$	0.9013	0.8611	0.9331	0.9481	0.9651
$MaxPool_{Graph}$	<b>0.9639</b>	<b>0.9451</b>	<b>0.9803</b>	<b>0.9862</b>	<b>0.9907</b>
模型	主关系 MRR	主关系 Hits@1	主关系 Hits@3	主关系 Hits@5	主关系 Hits@10
$1 \times 1 Conv$	0.9599	0.9389	0.9810	0.9844	0.9873
$LSTM$	0.8600	0.8037	0.9037	0.9255	0.9495
$MaxPool_{Graph}$	<b>0.9619</b>	<b>0.9423</b>	<b>0.9788</b>	<b>0.9851</b>	<b>0.9899</b>

表 4-5 三种聚合方法在 LENHGTN 模型的实体预测结果

模型	实体 MRR	实体 Hits@1	实体 Hits@3	实体 Hits@5	实体 Hits@10
$1 \times 1 Conv$	0.3706	0.2494	0.4376	0.5188	0.5972
$LSTM$	0.3385	0.2090	0.4010	0.5032	0.6412
$MaxPool_{Graph}$	<b>0.5071</b>	<b>0.3450</b>	<b>0.6034</b>	<b>0.7233</b>	<b>0.8182</b>
模型	主实体 MRR	主实体 Hits@1	主实体 Hits@3	主实体 Hits@5	主实体 Hits@10
$1 \times 1 Conv$	0.3777	0.2544	0.4481	0.5282	0.6047
$LSTM$	0.3402	0.2113	0.4137	0.5139	0.6534
$MaxPool_{Graph}$	<b>0.5186</b>	<b>0.3559</b>	<b>0.6162</b>	<b>0.7364</b>	<b>0.8322</b>

三种轻量化聚合方法的新增模型参数数量分别为:  $1 \times 1 Conv$  的新增参数数量与模型的网络层数  $L$  相同;  $LSTM$  通过采用矩阵分解与权值共享,新增参数数量为  $2NL$ ,与数据集的关系实体数量  $N$ 、模型的网络层数  $L$  正相关; 基于跨通道图最大池化

方法则没有新增模型参数。因此,首先从模型参数规模上,基于跨通道图最大池化的聚合方法 ( $MaxPool_{Graph}$ ) 更具有轻量化优势。

观察表 4-4、4-5 分别对关系、实体的预测结果,  $MaxPool_{Graph}$  聚合方法具有明显优势,且三种聚合方法的预测结果准确性之间整体上存在  $MaxPool_{Graph} > 1 \times 1Conv > LSTM$ ,其原因可能一方面在于原 NHGTN 模型复杂度已偏高的条件下,新增参数越多的聚合方法,给模型训练的压力越大,越难以训练模型得到较好的预测结果,所以有  $1 \times 1Conv > LSTM$ ,其中混合关系预测在 MRR 指标提升 6.8%。另一方法原因在于图变换网络的各层特征矩阵上,在同一位置的最大特征值确实对潜在节点预测具有关键影响,且远大于同一位置上的其他特征值,因此对于近似平均池化的  $1 \times 1Conv$  和  $LSTM$  两种方法,其在对同一位置不同层的特征值聚合时,实际效果等同于增加了噪声信息,削弱最大特征值的关键影响,从而降低了预测结果准确性,这一点在主实体预测上更为明显,  $MaxPool_{Graph}$  在指标 MRR 相比次优  $1 \times 1Conv$  方法提升了 37.3%。因此,上述实验结果证明了在 4.3.2 节分析思路的正确性,以及所提出的基于跨通道图最大池化的轻量化聚合方法的合理性与有效性。

注意到,两表中关系预测的效果提升明显小于实体预测,其原因在于数据集 JF17K 本身特点,其中实体与关系数量的比例约为 100:1,只有 322 个不同关系。因此,模型经过同样的训练,对关系的预测指标结果会更好,但同时其提升空间也会更小,使得在比较不同聚合方法的改进效果时,方法之间差异小,降低提出方法的说服力;而实体预测结果的提升空间远大于关系预测,使得所提出方法的改进结果明显高于其他方法,这在预测目标数量越大时,方法提升效果越明显,实验结果证明了所提出聚合方法的有效性。

### 3) LENHGTN 预测模型的有效性

为验证基于轻量化聚合的增强图变换链路预测模型相比现有方法的有效性与一定程度上的优越性。本节在 JF17K 数据集上与现有其他方法,包括基于主体三元组与辅助关系实体对表示的部分主流预测方法 (NeuInfer、HINGE)、基于张量分解 (S2S) 以及前面基于一组关系实体对表示的相关方法展开对比实验,结果如表 4-6、4-7 所示。需要说明的是,文献[10][2]对应模型 NeuInfer、HINGE 的关系预测指标不含 Hits@3,而除了少量模型能对关系预测外,其他模型实际缺少对关系的预测。

LENHGTN 模型在关系预测方面具有明显优势。在混合关系预测上,对比次优 tNaLP<sup>+</sup>模型在指标 MRR 和 Hits@1 分别提升了 0.83%与 0.76%;在主关系预测上,对比次优 HINGE 模型在指标 MRR 和 Hits@1 分别提升了 1.37%与 2.52%。说明 LENHGTN 模型能更加充分挖掘关系事实中的关系信息并对目标节点进行准确预

测。这是因为 LENHGTN 模型所基于的 $\mathcal{G}_{h-sta}$ 异构图表示充分建模了多元关系事实,相比 tNaLP<sup>+</sup>模型的一组关系实体对表示,突出了事实中三元组的结构主体性,而相比 NeuInfer、HINGE 又补全了辅助关系实体对之间的关联信息,并通过异构图结构对关系事实建立良好的表达同时补全了事实的异构特征。

表 4-6 LENHGTN 模型与部分现有主流模型的关系预测结果

模型	关系 MRR	关系 Hits@1	关系 Hits@3	关系 Hits@10
NeuInfer (2020)	0.861	0.832		0.904
tNaLP <sup>+</sup> (2021)	0.956	0.938	0.971	0.986
LENHGTN	<b>0.9639</b>	<b>0.9451</b>	<b>0.9803</b>	<b>0.9907</b>
模型	主关系 MRR	主关系 Hits@1	主关系 Hits@3	主关系 Hits@10
HINGE (2020)	0.9489	0.9191		0.9896
LENHGTN	<b>0.9619</b>	<b>0.9423</b>	<b>0.9788</b>	<b>0.9899</b>

表 4-7 LENHGTN 模型与部分现有主流模型的实体预测结果

模型	实体 MRR	实体 Hits@1	实体 Hits@3	实体 Hits@10
NeuInfer (2020)	0.517	0.436	0.553	0.675
S2S (2021)	0.528	0.457	0.570	0.690
tNaLP <sup>+</sup> (2021)	0.449	0.370	0.484	0.598
RAM (2021)	<b>0.539</b>	<b>0.463</b>	0.573	0.690
LENHGTN	0.5071	0.3450	<b>0.6034</b>	<b>0.8182</b>
模型	主实体 MRR	主实体 Hits@1	主实体 Hits@3	主实体 Hits@10
NeuInfer (2020)	0.431	0.342	0.466	0.611
HINGE (2020)	0.4554	<b>0.3677</b>	0.4679	0.6333
LENHGTN	<b>0.5186</b>	0.3559	<b>0.6162</b>	<b>0.8322</b>

在实体预测方面,LENHGTN 模型在部分指标上仍然具有一定的优势。在混合实体预测方面,模型在指标 Hits@3、Hits@10 优势较大,其中 Hits@10 相比次优 RAM、S2S 模型提升了 18.58%;在主实体预测方面,模型在指标 Hits@3、Hits@10 相比次优 HINGE 模型分别提升了 31.69%、31.41%。说明模型预测结果在一定范围内(如输出前 10 个候选预测结果),有较高的预测准确性,这主要在于模型能够更充分地建模和挖掘多元关系事实信息,这对一些输出结果的精确度要求不太高但又需要在一定范围内有较高准确度的应用,如智能推荐等,有较大意义。因此,通过上述对比实验,验证了 LENHGTN 模型的有效性与一定程度上的优越性

#### 4) 消融实验



为了验证本章提出的轻量化聚合模块的有效性，本节将 LENHGTN 模型与第三章的 NHGTN 模型进行了对比分析实验，结果如图 4-6 所示。

观察实验结果可知 LENHGTN 模型在关系实体预测上具有明显优势。模型在指标 MRR 和 Hits@1 上，主关系预测提升了 0.06%、0.16%，主实体预测提升了 11.72%、17.96%，其中主实体与实体的预测结果提升更为显著。LENHGTN 模型在预测结果的提升是因为轻量化聚合模块有效提取了图变换网络中分布在各层的关键特征信息，并对其聚合，输出的聚合元路径特征图相比 NHGTN 模型在输出层网络上的元路径特征图包含更多更准确的特征信息，从而得到较高的预测准确性，这在面向数量更多的实体目标预测上，结果提升更为显著。因此，通过消融实验，证明了所提出的轻量化聚合模块的有效性。

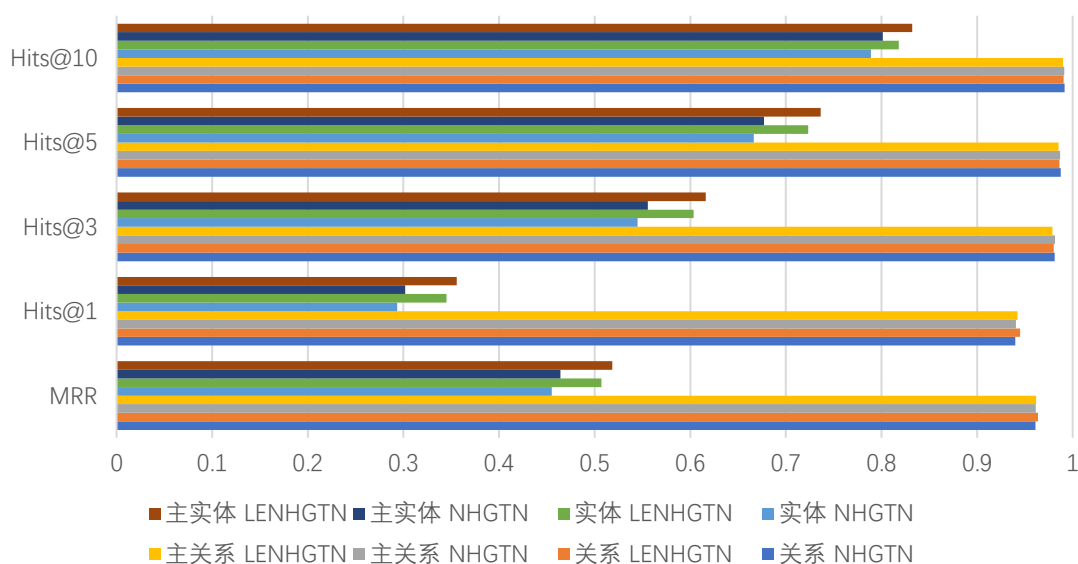


图 4-6 LENHGTN 模型与 NHGTN 模型的预测结果对比

## 4.5 本章小结

本章通过构建基于跨通道图最大池化的轻量化聚合模块，提出了基于轻量化聚合的增强图变换链路预测方法，用于解决图变换网络在深层网络上的信息丢失问题。首先，分别从定性和实验角度分析验证了深层网络信息丢失问题的产生原因及其存在，并简要说明了图变换链路预测模型复杂度较高的情况。其次，分析了现有基于主体三元组与辅助关系实体对表示的典型预测方法所存在的问题。然后，结合对图变换网络生成的元路径特征图的分析，提出了基于跨通道图最大池化的轻量化聚合方法，其通过网络通道对齐并利用图最大池化实现跨通道聚合分布在不同网络层上关键特征信息，进而提出了基于轻量化聚合的增强图变换链路预测模

型 LENHGTN，并详细介绍了模型结构和运行机制，最后在公开数据集 JF17K 上进行了多组分析对比验证实验，结果证明了本章所提出的轻量化聚合方法在解决图变换网络信息丢失问题上的有效性，也证明了本章提出的 LENHGTN 预测模型相对于现有其他模型在多元关系链路预测问题上的有效性和具有一定先进性。

## 第五章 多元关系事实预测系统设计与实现

公众人物信息具有公开性和广泛性，在现实中也易于获取。面向公众人物信息的查询系统在网络宣传和构建人物画像方面都具有一定现实意义。在本章将实现一个面向公众人物信息查询的多元关系链路预测系统，根据用户输入的相关事实信息，对其中的待测关系或者实体目标进行预测，输出预测结果及其排名，从而再次验证第三章和第四章所提出的多元关系链路预测模型的可行性。本章后续小节将分别从需求分析、总体设计、系统实现和系统展示四个部分展开详细介绍。

### 5.1 需求分析

本节通过对待开发的面向公众人物信息查询的多元关系链路预测系统进行需求分析，从中总结出系统需要涵盖的主要功能，并设计系统整体的数据流向。该系统用例分析如图 5-1 所示，主要包括实体预测、关系预测、数据管理、日志管理等。

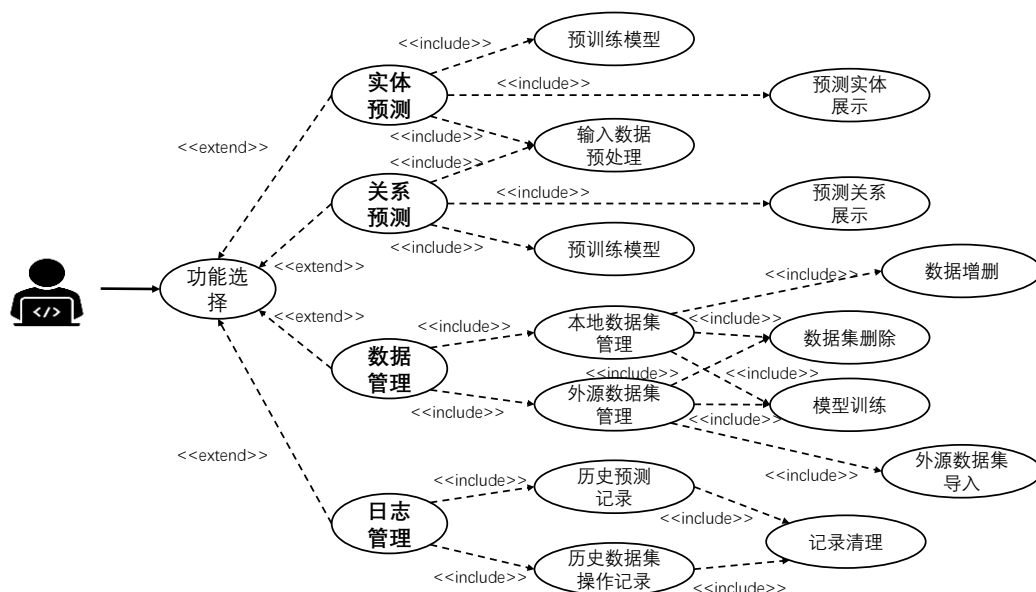


图 5-1 系统用例图

系统整体数据流向如图 5-2 所示，主要分为关系、实体预测事务和数据集处理事务。针对关系、实体预测事务，系统控制首先将接收的用户数据输入到数据预处理模块进行处理，主要是对用户数据的编码映射、数据字典校验，得到由事实关系实体索引组成的嵌入矩阵；同时，系统将用户选用数据集、预测对象类型以及对应系统参数输入预测模型模块实现预测模型选用。然后利用预测模型对输入的关系

事实中待预测目标进行预测，结果输入到前端展示，并将本次预测事件及预测结果简要记录到日志中。针对数据集处理事务，通过系统控制首先将用户输入数据集信息和相关系统信息输入数据集管理模块，主要操作包括本地数据集数据增删、外源数据集上传和导入本地；同时系统将相关模型参数信息输入到预测模型模块进行对应预测模型选择。然后数据集模块将目标数据集信息输入预测模型，对目标模型进行训练，并在训练结束时将本次数据集处理事件记录日志。

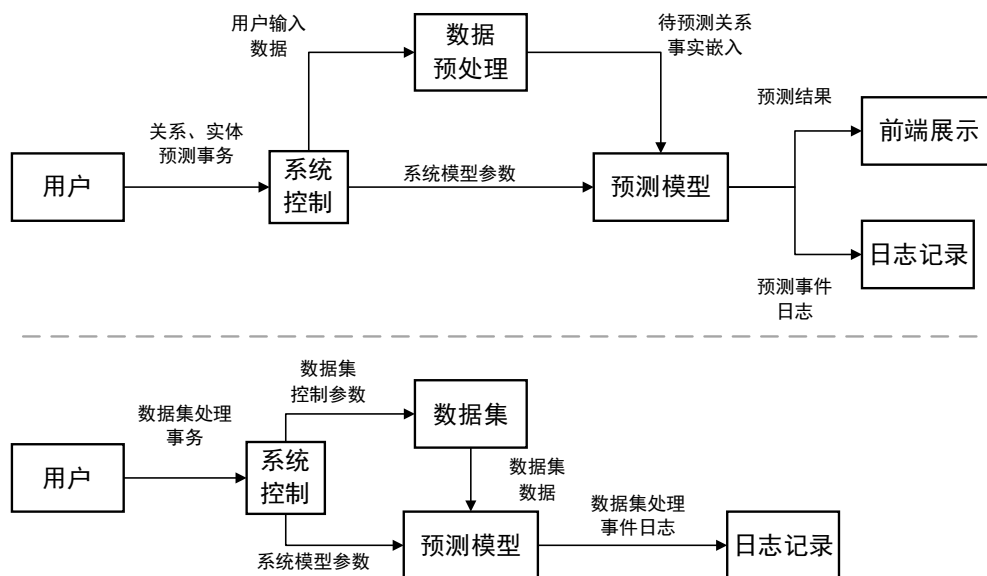


图 5-2 系统数据流图

基于上面用例及数据流向图的分析，本章开发的多元关系链路预测系统的主要功能包括：

- (1) 用户输入信息的预处理
- (2) 目标关系事实的实体预测
- (3) 目标关系事实的关系预测
- (4) 目标预测结果可视化展示
- (5) 本地数据集处理
- (6) 外源数据集导入
- (7) 日志记录

## 5.2 总体设计

系统整体采用 B/S（浏览器/服务器）架构，后端采用 PyCharm 集成开发环境并以 Python 作为开发语言，前端页面设计基于 HTML 和 CSS3，数据存储分别采

用关系型数据库 MySQL 存储数据库控制信息和非关系型数据库 MongoDB 存储数据库数据信息、日志信息。

多元关系链路预测系统整体架构如图 5-3 所示,

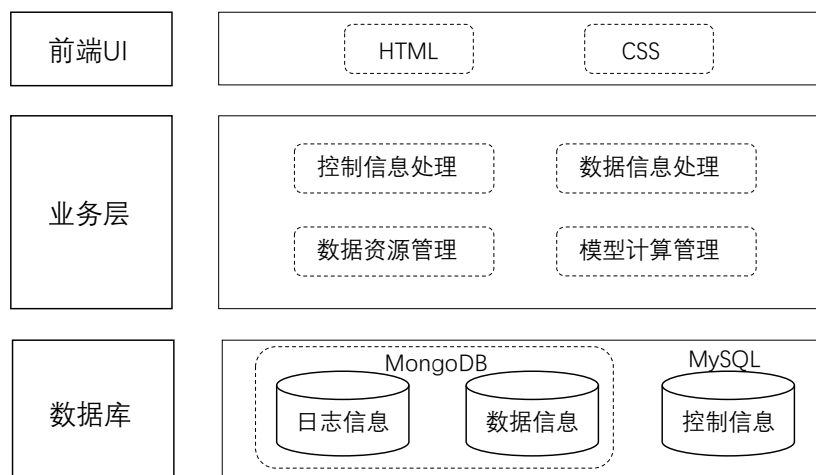


图 5-3 系统整体架构图

最上层是前端 UI，主要基于 HTML5、CSS3 进行系统界面展示和功能集成。业务层作为系统核心，包括控制信息处理、数据信息处理、数据资源管理和模型计算管理，为多元关系链路预测系统后台信息及资源流向进行管理。数据库中主要利用 MySQL 存储相关控制信息和 MongoDB 存储数据集信息和日志信息。

系统一共由四个模块构成，如图 5-4 所示，

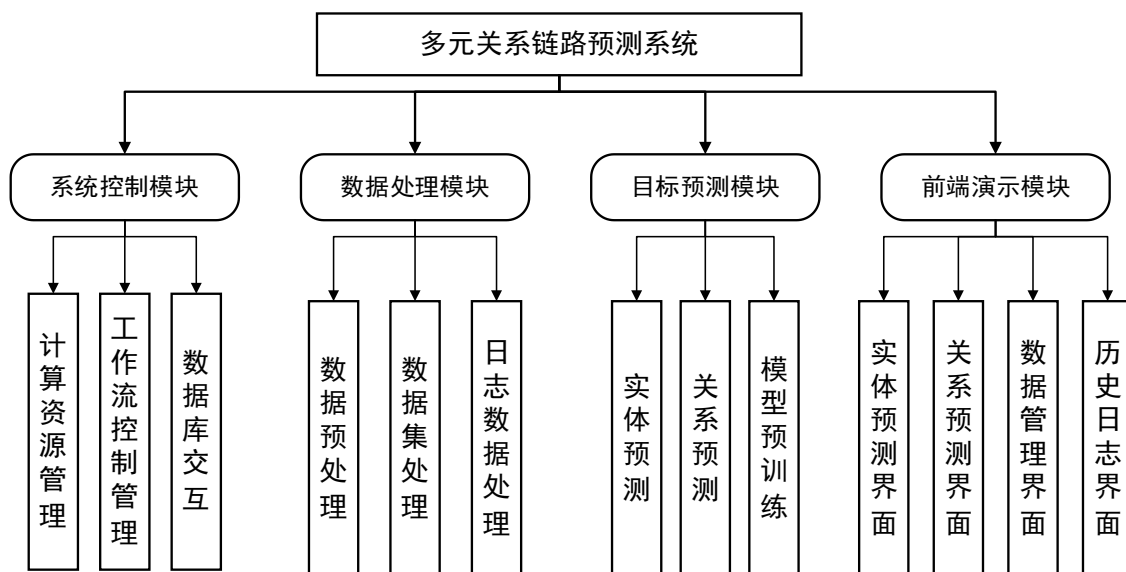


图 5-4 系统模块分解图

(1) 系统控制模块: 该模块包括计算资源管理、工作流控制管理和数据库交互三个子模块, 主要功能是对系统 GPU 算力实现检测和调配, 保证系统整个生命周期的完整运行, 以及完成与数据库的交互工作。

(2) 数据处理模块: 该模块主要由数据预处理、数据集处理、日志数据处理三个子模块构成, 主要用于数据流向控制、完成用户输入数据的统一预处理、本地及外源数据集上的处理、以及事件日志数据存储等工作。

(3) 目标预测模块: 该模块包括实体预测、关系预测、模型预训练, 主要用于实现对用户输入关系事实中目标对象的预测, 以及在数据集发生变化后调整模型参数信息并对模型进行训练更新或者新建模型训练副本。

(4) 前端展示模块: 该模块包括实体预测界面、关系预测界面、数据管理界面、历史日志界面, 主要实现与用户的交互以及结果输出展示。

## 5.3 系统实现

本节主要对多元关系链路预测系统中各个模块的具体实现展开详细介绍。

### 5.3.1 系统控制模块

系统控制模块连接其他各子模块入口, 保证系统生命周期的完整运行。系统整体工作流程如图 5-5 所示。

系统控制模块对实体、关系预测, 数据管理, 历史记录管理四个不同的事务进行管理。对于实体、关系预测事务, 用户首先选择预测类型, 输入待预测关系事实, 确定预测对象, 然后经过调用后台数据预处理和相应预测模型对目标进行预测, 最后将预测结果输出到前端展示, 同时将本次预测事件记录到后台日志。对于数据管理事务, 用户可选择对本地数据进行修改或者是导入外源数据集, 选择本地数据修改, 则先输入需要修改的关系事实样本, 经过数据预处理将其转化为数据集统一格式添加到数据集中, 再对需要删除的原样本进行删除, 数据集更新后还需对相应模型进行重新训练; 选择外源数据集导入则需用户根据说明自行上传数据集, 然后建立模型副本并调整模型参数进行训练, 完成模型训练后将本次事件处理记录到后台日志。对于历史记录事务, 用户可查询历史日志, 根据需要删除历史预测记录并更新相应日志信息。

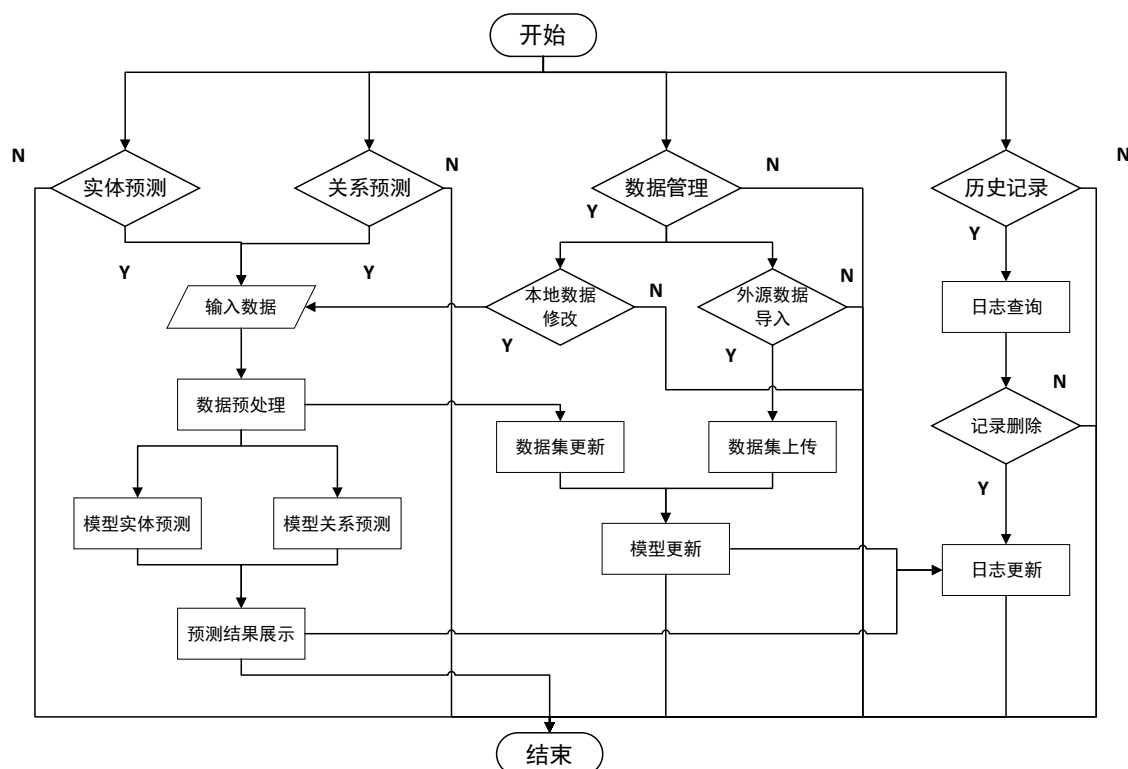


图 5-5 系统工作流程图

### 5.3.2 数据处理模块

数据处理模块包含用户输入数据的预处理、数据集的数据交互以及日志信息记录和更新。其中数据预处理是针对用户输入的关系事实样本，将样本中的实体关系通过映射转化为系统数据集中样本统一格式，并将转化后的关系实体与数据集对应的数据字典进行校验，检测是否为数据集已知对象，只有全部检测为已知才能继续进行后续预测操作。但校验过程在对数据集增加新关系事实样本时不启用，只需完成数据映射转化。数据集处理主要是对数据集中的数据交互，控制本地数据集新增样本数据或者删除样本数据，控制外源数据集上传以及本地存储。日志数据处理主要是对日志数据的新增、删除处理。

### 5.3.3 目标预测模块

目标预测模块的核心是本文第三章、第四章提出的多元关系链路预测模型。对于实体、关系预测，根据用户选择的目标数据集，将预处理输出的待预测关系事实输入到对应已训练好的预测模型，得到预测结果并将其解码传输到前端进行展示。模型预训练是针对具体更新后的数据集或者新导入的外源数据集，在完成数据更新或数据导入后，选择启动模型预训练，并在训练完成后提示可进行目标预测。

### 5.3.4 前端演示模块

前端演示模块主要包括实体与关系预测界面、数据管理界面和历史日志界面，主要构建系统前端展示框架，网页基本采用左右布局格式，左边是功能导航栏，右边为对应功能的具体展示。系统采用的基础 UI 布局如图 5-6 所示，其中每个界面实际视图及功能将在下面 5.4 节展示。

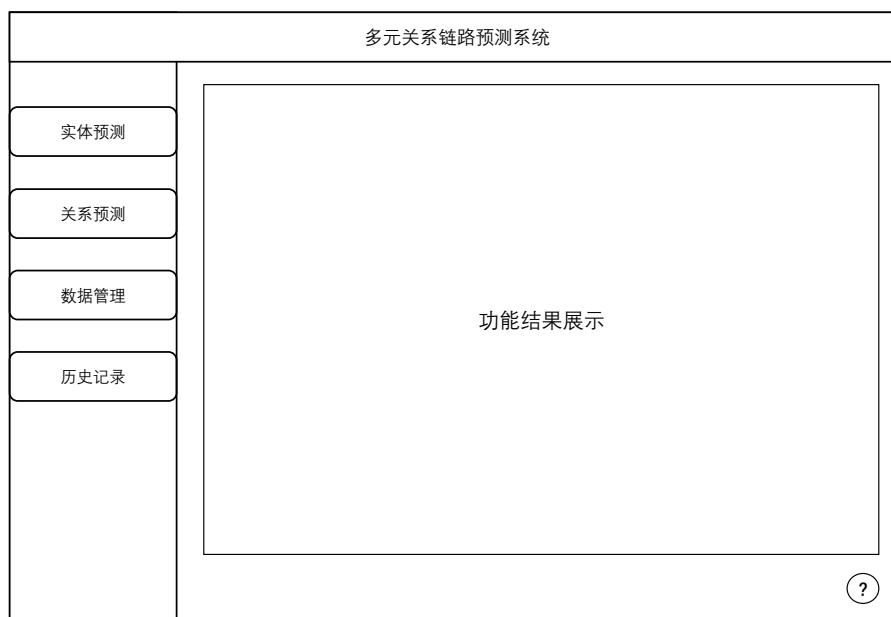


图 5-6 系统 UI 原型图

## 5.4 系统展示

本节将主要展示该系统四个功能模块（实体预测、关系预测、数据管理和历史记录）对应的具体界面，及其操作流程。系统初始界面如图 5-7 所示。

### 5.4.1 实体预测界面

实体预测界面如图 5-8 所示，在最上方的文本框内输入待预测的关系事实文本，系统将自动将其分解到在下方的头尾实体、关系及辅助关系实体对选框中，用户可以手动对选框内的对象进行修改，并输入样本目标数据集，系统会自动将用户输入数据与系统内部数据集名称、数据集对应关系实体字典进行校验，如果输入数据对象存在则“√”否则为“×”，只有校验成功的待测样本才能完成目标预测过程。系统将根据输入的数据集选择对应的实体预测模型，输出结果将以（排名，预测对象名称）的形式在下方文本框中展示。系统支持最多进行两个样本的同时预测。



### 5.4.2 关系预测界面

关系预测界面如图 5-9 所示，与上面实体预测界面和功能操作保持一致，用户在最上方文本框内输入待预测关系事实文本，系统将自动将其分解或用户手动输入到对应选框，由系统完成校验后，完成目标预测过程。预测结果以（排名，预测对象名称）的形式输出在下方文本框中展示。系统支持最多进行两个样本的同时预测。

### 5.4.3 数据管理界面

数据管理界面如图 5-10 所示，可对系统中已有数据集进行删除，以及对本地数据集进行数据处理和外源数据集导入。

对于本地数据集可进行指定关系事实样本的删除，操作如图 5-11 所示，注意为了避免破坏数据集，只支持对少数极个别样本的删除，以及添加新的关系事实样本到该数据集中。其中删除指定关系事实样本需要用户输入数据集名称和相关实体和关系信息，系统执行与前面实体关系预测相同的数据预处理和校验过程，启动查找后，系统将在该数据集中查找包含用户输入信息的全部相关关系事实样本并输出，用户可选择样本进行删除，之后需要用户点击“重新训练模型”按钮，对模型进行更新。新增关系事实是由用户输入新增关系事实样本，然后点击“重新训练模型”，系统自动将新增样本添加到该数据集并对模型进行更新。

外源数据集导入需要根据系统说明点击上传按钮，将相关训练集、验证集、测试集和数据集相关说明文件进行上传，如图 5-12 所示，其中数据文件仅支持 json 格式，说明文件仅支持 txt 格式。上传后，用户可自行选择是否将数据集添加到本地存储，未选择添加到本地，该数据集将在固定一段时间后自动删除。上传完数据集后需用户点击启动模型训练，为该数据集训练出对应模型副本。

### 5.4.4 历史记录界面

历史记录界面如图 5-13 所示，用户可查阅其中历史预测记录和历史数据集操作记录，根据需要用户可对历史预测记录进行删除。



图 5-7 系统初始界面演示界面



图 5-8 系统实体预测功能演示界面

多元关系链路预测系统

实体预测

关系预测

数据管理

历史记录

The 2021 Nobel Prize in Chemistry [Q] German scientist Benjamin List and Scotland-born scientist David W.C. MacMillan of Princeton University, the Royal Swedish Academy of Sciences announced on Wednesday.

头实体

Nobel Prize in Chemistry

主关系

尾实体

Benjamin List

数据集

JF17K

辅助关系实体对

(point-in-time, 2021),(nationality-of, German),(together-with, David W.C. MacMillan),(nationality-of, Scotland-born),(organization-of, Princeton)

[[1, awarded-to), (2, award-nominated), (3, award-received), (4, receive), (5, award-for), (6, announced-to), (7, announced-by), (8, made-by), (9, sent-to), (10, honored-for)]

开始预测

请输入关系事实文本 (待预测对象请用[Q]标记)

头实体

主关系

尾实体

数据集

辅助关系实体对

开始预测

< 1 2 3 4 5 >

?

图 5-9 系统关系预测功能演示界面

多元关系链路预测系统

实体预测

关系预测

数据管理

历史记录

本地 外源

类型	名称	数据文件	操作
<input checked="" type="checkbox"/>	本地	JF17K	all.json, test.json, train+valid.json, train.json, valid.json, vocab.txt, readme-jf17k.txt
<input type="checkbox"/>	本地	WD50K_100	all.json, test.json, train+valid.json, train.json, valid.json, vocab.txt, readme-wd50k100.txt
<input type="checkbox"/>	外源	Wikipedea-	all.json, test.json, train+valid.json, train.json, valid.json, vocab.txt, readme-wikipedia.txt

< 1 2 3 4 5 >

?

图 5-10 系统数据管理功能演示界面 1

59

中国知网 <https://www.cnki.net>

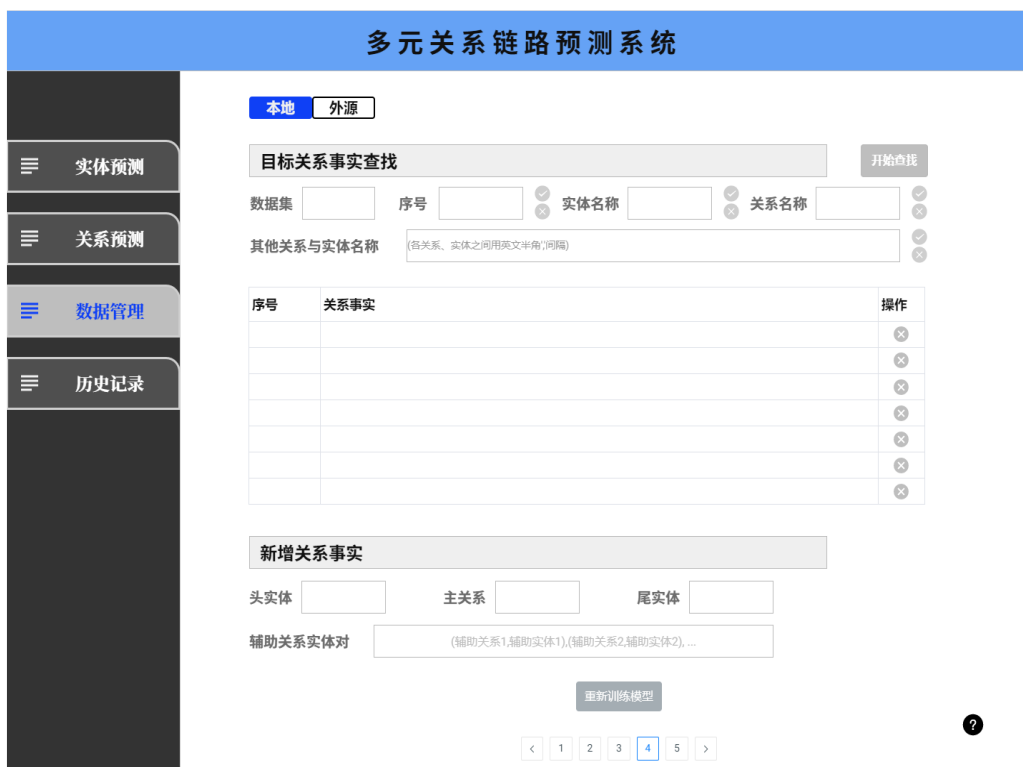


图 5-11 系统数据管理功能演示界面 2



图 5-12 系统数据管理功能演示界面 3



图 5-13 系统历史记录功能演示界面

## 5.5 系统测试

为了保证本系统能够正常稳定运行，本节对系统各功能模块分别进行测试。

人机交互测试，目标是测试系统前端各功能模块可用（正常切换、可输入等操作）、用户体验流畅、用户可得到系统及时反馈。测试用例如表 5-1 所示，

表 5-1 人机交互测试用例

编号	测试名称	测试过程	预期结果
1	前端功能界面切换	随机点击四个测试功能选项进行切换	各选线及子选项之间切换流畅
2	实体预测界面输入	在实体预测界面按位置输入测试样例	有输出返回结果
3	关系预测界面输入	在关系预测界面按位置输入测试样例	有输出返回结果
4	数据管理本地界面输入	在数据管理/本地界面上输入测试样例	有输出返回结果
5	数据管理外源界面上上传	在数据管理/外源界面输入上传数据集名称并点击上传	可成功打开文件夹
6	历史记录查询	在历史记录界面相关操作记录，任一选择预测记录删除	若成功删除预测记录则返回提示

实体预测功能测试，目标是测试系统能够对多元关系事实进行预测、并输出正确预测结果。测试用例如表 5-2 所示，

表 5-2 实体预测测试用例

编号	测试名称	测试过程	预期结果
1	关系、实体目标识别	在文本框内输入多元关系事实样例，系统自动识别抽取事实中的关系实体信息并在对应框图内输出	框内正确输出事实中的实体、关系
2	目标实体预测	修改并确认框图内已知实体关系、待测实体等信息无误后，点击“开始预测”按钮	输出框内返回前 top10 的实体预测结果及排名（包含正确结果）

关系预测功能测试，目标是测试系统能够对多元关系事实进行预测、并输出正确预测结果。测试用例如表 5-3 所示，

表 5-3 关系预测测试用例

编号	测试名称	测试过程	预期结果
1	关系、实体目标识别	在文本框内输入多元关系事实样例，系统自动识别抽取事实中的关系实体信息并在对应框图内输出	框内正确输出事实中的实体、关系
2	目标关系预测	修改并确认框图内已知实体关系、待测关系等信息无误后，点击“开始预测”按钮	输出框内返回前 top10 的关系预测结果及排名（包含正确结果）

数据管理功能测试，目标是测试系统能够对多元关系事实数据、数据集进行增删查操作。测试用例如表 5-4 所示，

表 5-4 数据管理测试用例

编号	测试名称	测试过程	预期结果
1	多元关系事实增删查	在数据管理/本地界面按位置输入测试样例，点击“开始查找”；然后在返回结果中删除该事实；最后在新增关系事实中按位置输入测试样例，并点击“重新训练模型”	正确查找并输出关系事实；删除后返回成功提示
2	多元关系数据集增删	在数据管理界面随机选择一项数据集，点击删除按钮；在数据管理/外源界面点击上传按钮，选择之前删除的数据集进行上传；可任选“添加到本地”；最后点击“自动模型训练”	数据集删除后返回成功提示；新增数据集上传后返回成功提示，若选择了添加到本地，则其类型为“本地”否则“外源”

## 5.6 本章小结

本章主要介绍了面向公众人物信息查询的多元关系链路预测系统。从需求分析、总体设计和系统实现三个方面对该系统整体开发流程进行了详细分析与介绍,然后通过可视化界面对系统的实体预测、关系预测、数据管理和历史日志等核心功能进行效果演示,并给出相关测试用例。通过展示本系统操作和预测结果,证明了本文提出的面向多元关系的链路预测方法具有一定的实用性。

## 第六章 总结与展望

当今网络世界中数据信息总量呈现爆炸式增长,利用知识图谱和数据之间关联加速信息搜索、提高查询结果准确率成为智能检索中一种常见技术手段。此外,知识图谱还广泛应用于智能问答、智能推荐等新型人工智能领域,为应用提供更加准确快速的服务,对于推动人工智能的发展具有重要意义。多元关系的链路预测是对包含多个实体与一个或多个关系构成的复杂关系事实的预测,其能够直接利用知识图谱中的复杂关系事实进行学习,减少了信息丢失与知识歧义,从而提高了预测结果的准确性。针对现有方法中存在对多元关系事实理解不充分、信息利用不足的问题,本文从多元关系事实本身出发,提出了一种多元关系的异构图表示形式,用于充分建模关系事实中实体关系之间的异构交互,并结合图变换网络提出了基于多元关系异构图表示的图变换网络链路预测模型。进一步针对图变换深层网络中的信息丢失问题,构建了基于跨通道图最大池化的轻量化聚合模块,对图变换网络预测模型进行了增强改进,解决了信息丢失问题。模型增强了对多元关系事实的理解,实现了对关系事实内在信息的充分挖掘、提高了预测准确性。

### 6.1 全文总结

面向多元关系链路预测任务,本文提出了一种基于异构图表示的图变换网络链路预测方法和基于轻量化聚合的增强图变换链路预测方法,主要工作如下:

(1) 调研并分析了多元关系链路预测领域的研究方法以及相关进展。通过对国内外相关论文的调研,梳理了多元关系链路预测领域的发展脉络及现阶段研究状况,并将该领域所提出的现有方法模型进行了粗略划分。

(2) 探讨分析了关于多元关系事实的两种基本表示形式,并分别对其表示形式所存在的潜在问题进行了分析。然后在此基础上,提出构建了多元关系的异构图表示形式,并将其作为后续所提出模型的基础内容。

(3) 提出了一种基于多元关系异构图表示的图变换网络预测方法。方法基于异构图表示的关系事实,结合图变换网络、图采样、以及图降噪模块,将图变换网络应用到大规模多元关系数据集上,实现了对多元关系事实中实体和关系的预测。

(4) 提出了一种基于轻量化聚合的增强图变换链路预测方法。针对图变换网络在深层网络上存在信息丢失的问题,本文通过理论分析和实验分别进行了验证,并在此基础上提出了一种基于跨通道图最大池化的轻量化聚合方法,并对图变换网



络链路预测模型进行了改进,有效提取和聚合了图变换网络各层上关键特征,解决了深层网络信息丢失问题。

(5) 设计并构建了面向公众人物信息查询的多元关系链路预测系统。

本文的主要创新点如下:

(1) 针对现有方法对于多元关系事实理解不充分,信息利用不足,本文提出了一种基于多元关系异构图表示的图变换网络链路预测方法。利用异构图表示充分建模关系事实中的异构交互信息,然后引入了一种图变换神经网络并利用图采样和降噪使其适用于大规模多元关系数据集,实现了对多元关系事实中任意实体和关系的预测。

(2) 针对图变换深层网络上存在信息丢失问题,结合当前模型的复杂度较高的情况,本文提出了基于轻量化聚合的增强图变换链路预测方法。通过分析图变换网络各层输出元路径特征图上最大特征值对应关键特征,提出了基于跨通道图最大池化的轻量化聚合方法,实现了对模型各层网络上关键特征的有效提取与聚合,补全了网络丢失信息,从而提升了预测结果的准确性。

## 6.2 未来展望

本文在多元关系链路预测问题上提出了一种新的基于异构图表示形式并结合图变换网络的预测方法,在大规模多元关系数据集上验证了方法的可行性和有效性,相较其他现有方法具有一定的先进性。然而,在研究和实验过程当中发现所提出的方法依旧存在进一步的提升空间。

(1) 优化模型架构,降低模型在大规模数据集训练难度。

当前方法受到数据集规模和图变换网络复杂度偏高的影响,在大规模数据集上存在训练时间过长,模型训练难度较大的问题。未来可以适当优化模型,如改进图采样模块,并建立大规模数据样本采样策略,从而降低模型的训练难度,缩短训练时间。

(2) 改进关系实体嵌入,增大相近关系事实之间的差异性。

现实知识图谱中存在一些关系事实,它们有极为相似的关系实体,如一组相近关系事实之间只有 1~2 个实体或者关系的不同,且不同之处都位于关系事实的同一位置,这种情况极大影响了本文方法的预测效果,因此未来的改进重点在于针对这类相近关系事实,建立一种糅合异构信息的关系实体嵌入表示策略以及相应模型用于预训练,实现对其中关系实体节点嵌入的最大程度区分,从而有效提高预测结果的准确性。

## 参考文献

- [1] Rossi A, Barbosa D, Firmani D, et al. Knowledge graph embedding for link prediction: a comparative analysis[J/OL]. ACM Transactions on Knowledge Discovery from Data. 2021, 15(2): 14:1-49, <https://export.arxiv.org/pdf/2002.00819>.
- [2] Rosso P, Yang D, Cudré-Mauroux P. Beyond triplets: hyper-relational knowledge graph embedding for link prediction[C]. Proceedings of The Web Conference 2020, Taipei, Taiwan, 2020: 1885-1896.
- [3] Guan S, Jin X, Wang Y, et al. Link prediction on n-ary relational data[C]. The World Wide Web Conference 2019, San Francisco, CA, USA, 2019: 583-593.
- [4] Guan S, Jin X, Guo J, et al. Link prediction on n-ary relational data based on relatedness evaluation[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(1): 672-685.
- [5] Galkin M, Trivedi P, Maheshwari G, et al. Message passing for hyper-relational knowledge graphs[C]. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 2020:7346-7359.
- [6] Yu D, Yang Y. Improving hyper-relational knowledge graph completion[OL]. 2021, <https://doi.org/10.48550/arXiv.2104.08167>.
- [7] Liu Y, Yao Q, Li Y. Role-aware modeling for n-ary relational knowledge bases[C]. In Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 2021: 2660-2671
- [8] Wang Q, Wang H, Lyu Y, et al. Link prediction on n-ary relational facts: a graph-based approach[C]. In Findings of the Association for Computational Linguistics: ACL-IJCNLP, Online, 2021: 396-407
- [9] Wen J, Li J, Mao Y, et al. On the representation and embedding of knowledge bases beyond binary relations[C/OL]. International Joint Conference on Artificial Intelligence, 2016:1300-1307, <https://arxiv.org/pdf/1604.08642.pdf>.
- [10] Guan S, Jin X, Guo J, et al. NeuInfer: Knowledge inference on n-ary facts[C]. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Online, 2020: 6141-6151.

- [11] Zhang R, Li J, Mei J, et al. Scalable instance reconstruction in knowledge bases via relatedness affiliated embedding[C]. In Proceedings of the 2018 World Wide Web Conference (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018: 1185-1194.
- [12] Yan S, Zhang Z, Sun X, et al. HYPER<sup>2</sup>: Hyperbolic embedding for hyper-relational link prediction[J]. Neurocomputing, 2022, 492: 440-451.
- [13] Di S, Yao Q, Chen L. Searching to sparsify tensor decomposition for n-ary relational data[C]. In Proceedings of the Web Conference 2021 (WWW '21). Association for Computing Machinery, New York, NY, USA, 2021: 4043-4054.
- [14] Brayne A, Wiatrak M, Corneil D. On masked language models for contextual link prediction[C]. In Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures, Dublin, Ireland and Online. Association for Computational Linguistics, 2022: 87-99.
- [15] Zhang Y, Xu H, Zhang X, et al. TRFR: A ternary relation link prediction framework on knowledge graphs[J]. Ad Hoc Networks, 2021, 113: 102402.
- [16] Hamilton W, Ying Z, Leskovec J. Inductive representation learning on large graphs[C]. Advances in Neural Information Processing Systems 30 (NIPS 2017), Long Beach, CA, USA, 2017: 1024-1034.
- [17] Zeng H, Zhou H, Srivastava A, et al. GraphSAINT: Graph sampling based inductive learning method[C/OL]. ICLR 2020, <https://arxiv.org/pdf/1907.04931.pdf>.
- [18] Zou D, Hu Z, Wang Y, et al. Layer-dependent importance sampling for training deep and large graph convolutional networks[C]. 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada. 2019: 11247-11256.
- [19] Chen J, Ma T, Xiao C. FastGCN: Fast learning with graph convolutional networks via importance sampling[C/OL]. In 6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, 2018, <https://arxiv.org/pdf/1801.10247.pdf>.
- [20] Yun S, Jeong M, Kim R, et al. Graph transformer networks[C]. Advances in Neural Information Processing Systems 32 (NeurIPS 2019), 2019: 11960-11970.
- [21] Yun S, Jeong M, Yoo S, et al. Graph transformer networks: learning meta-path graphs to improve GNNs[J]. Neural Networks: the official journal of the International Neural Network Society, 2021, 153: 104-119.

- [22] Hu Z, Dong Y, Wang K, et al. Heterogeneous graph transformer[C]. In Proceedings of The Web Conference 2020 (WWW '20). Association for Computing Machinery, New York, NY, USA, 2020: 2704-2710.
- [23] Zhang C, Song D, Huang C, et al. Heterogeneous graph neural network[C]. In Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD '19). Association for Computing Machinery, New York, NY, USA, 2019: 793-803.
- [24] Wang X, Ji H, Shi C, et al. Heterogeneous graph attention network[C]. In The World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 2019: 2022-2032.
- [25] Huang J, Liu X, Song Y. Hyper-path-based representation learning for hyper-networks[C]. In Proceedings of the 28th ACM International Conference on Information and Knowledge Management (CIKM '19). Association for Computing Machinery, New York, NY, USA, 2019: 449-458.
- [26] Wang W, Zhang W, Liu S, et al. Incorporating link prediction into multi-relational item graph modeling for session-based recommendation[J]. IEEE Transactions on Knowledge and Data Engineering, 2021, 35(3): 2683-2696.
- [27] Liu X, Yan M, Deng L, et al. Sampling methods for efficient training of graph convolutional networks: a survey[J]. IEEE/CAA Journal of Automatica Sinica, 2022, 9(2): 205-234.
- [28] Fan X, Gong M, Tang Z, et al. Deep neural message passing with hierarchical layer aggregation and neighbor normalization[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(12): 7172-7184.
- [29] Huang G, Liu Z, Van Der Maaten L, et al. Densely connected convolutional networks[C], 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017: 2261-2269.
- [30] Yu F, Wang D, Shelhamer E, et al. Deep layer aggregation[C]. Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2018: 2403-2412.
- [31] Lin T, Dollár P, Girshick R, et al. Feature pyramid networks for object detection[C], 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017: 936-944.
- [32] Zhao J, Fang Y, Li G. Recurrence along depth: deep convolutional neural networks with recurrent layer aggregation[C]. Advances in Neural Information Processing Systems 34 (NeurIPS 2021): 10627-10640.

- [33] Wang Q, Huang P, Wang H, et al. CoKE: Contextualized knowledge graph embedding[OL]. 2019, <https://arxiv.org/pdf/1911.02168.pdf>.
- [34] Allen C, Balazevic I, Hospedales T, et al. Interpreting knowledge graph relation representation from word embeddings[C/OL]. International Conference on Learning Representations 2021, [https://www.pure.ed.ac.uk/ws/portalfiles/portal/194766242/Interpreting\\_Knowledge\\_ALLEN\\_DOA12012021\\_AFV.pdf](https://www.pure.ed.ac.uk/ws/portalfiles/portal/194766242/Interpreting_Knowledge_ALLEN_DOA12012021_AFV.pdf).
- [35] Zhao J, Wang X, Shi C, et al. Heterogeneous graph structure learning for graph neural networks[C]. Proceedings of the AAAI Conference on Artificial Intelligence, 2021, 35(5): 4697-4705.
- [36] Gilmer J, Schoenholz S, Riley P, et al. Neural message passing for quantum chemistry[C]. Proceedings of the 34th International Conference on Machine Learning, PMLR, 2017: 1263-1272.
- [37] Chen J, Zheng S, Song Y, et al. Learning attributed graph representations with communicative message passing transformer[OL]. 2021, <https://arxiv.org/abs/2107.08773>.
- [38] Wang H, Ren H, Leskovec J. Relational message passing for knowledge graph completion[C]. In Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining (KDD '21). Association for Computing Machinery, New York, NY, USA, 2021: 1697-1707.
- [39] Trouillon T, Welbl J, Riedel S, et al. Complex embeddings for simple link prediction[C]. Proceedings of the 33rd International Conference on Machine Learning, PMLR, 2016, 48: 2071-2080.
- [40] Grattarola D, Zambon D, Bianchi F, et al. Understanding pooling in graph neural networks[J/OL]. IEEE Transactions on Neural Networks and Learning Systems, 2022: 1-11, <https://arxiv.org/abs/2110.05292>.
- [41] Lecun Y, Bottou L, Bengio Y, et al. Gradient-based learning applied to document recognition[J], in Proceedings of the IEEE, 1998, 86(11): 2278-2324.
- [42] Ying Z, You J, Morris C, et al. Hierarchical graph representation learning with differentiable pooling[C]. In Advances in Neural Information Processing Systems, NeurIPS, 2018: 4805-4815.
- [43] Liu C, Zhan Y, Li C, et al. Graph pooling for graph neural networks: progress, challenges, and opportunities[OL]. 2022, <https://arxiv.org/abs/2204.07321>.
- [44] Dettmers T, Minervini P, Stenetorp P, et al. Convolutional 2D knowledge graph embeddings[C]. AAAI Conference on Artificial Intelligence 2018: 1811-1818.

- [45] Nguyen D, Nguyen T, Nguyen D, et al. A novel embedding model for knowledge base completion based on convolutional neural network[C]. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana. 2018, 2: 327-333.
- [46] Gui H, Liu J, Tao F, et al. Embedding learning with events in heterogeneous information networks[J], IEEE Transactions on Knowledge and Data Engineering, 2017, 29(11): 2428-2441.
- [47] Gui H, Liu J, Tao F, et al. Large-scale embedding learning in heterogeneous event data[C], IEEE 16th International Conference on Data Mining (ICDM), Barcelona, Spain, 2016: 907-912.
- [48] Scarselli F, Gori M, Tsoi A, et al. The graph neural network model[J], IEEE Transactions on Neural Networks, 2009, 20(1): 61-80.
- [49] Zamini M, Reza H, Rabiei M. A review of knowledge graph completion[J]. Information. 2022; 13(8): 396.
- [50] Ali M, Berrendorf M, Galkin M, et al. Improving inductive link prediction using hyper-relational facts[C]. The Semantic Web - ISWC 2021 - 20th International Semantic Web Conference, ISWC 2021, 12922: 74-92.
- [51] Wu L, Cui P, Pei J, et al. Graph neural networks: foundations, frontiers, and applications[M]. Springer, Singapore, 2022: 0-725.
- [52] Ji S, Pan S, Cambria E, et al. A survey on knowledge graphs: representation, acquisition, and applications[J]. IEEE Transactions on Neural Networks and Learning Systems, 2022, 33(2): 494-514.
- [53] Wang Q, Mao Z, Wang B, et al. Knowledge graph embedding: a survey of approaches and applications[J]. IEEE Transactions on Knowledge and Data Engineering, 2017, 29(12): 2724-2743.
- [54] Kipf T, Welling M. Semi-supervised classification with graph convolutional networks[C/OL]. 5th International Conference on Learning Representations, ICLR 2017, Toulon, France, 2017, <https://arxiv.org/abs/1609.02907>.
- [55] 刘峤,李杨,段宏等.知识图谱构建技术综述[J].计算机研究与发展, 2016, 53(03): 582-600.
- [56] Vrandečić D, Krötzsch M. Wikidata: a free collaborative knowledgebase[J]. Communications of the ACM, 2014, 57(10): 78-85.
- [57] Auer S, Bizer C, Kobilarov G, et al. Dbpedia: a nucleus for a web of open data[C]. The Semantic Web, 6th International Semantic Web Conference, 2nd Asian Semantic Web Conference, ISWC 2007 + ASWC 2007, Busan, Korea, 2007: 722-735.

- [58] Suchanek F, Kasneci G, Weikum G. Yago: a core of semantic knowledge[C]. In Proceedings of the 16th international conference on World Wide Web (WWW '07). Association for Computing Machinery, New York, NY, USA, 2007: 697–706.
- [59] Liu H, Singh P. ConceptNet - a practical commonsense reasoning tool-kit[J]. BT Technology Journal, 2004, 22(4): 211-226.
- [60] Chen L, Mao X, Zhang Y, et al. An efficient ROS package searching approach powered by knowledge graph[C]. The 33rd International Conference on Software Engineering and Knowledge Engineering, SEKE, 2021: 411-416.
- [61] Liu F, Li T. An efficient application searching approach based on user review knowledge graph[C]. The 32nd International Conference on Software Engineering and Knowledge Engineering, SEKE, 2020: 406-412.
- [62] Rudnik C, Ehrhart T, Ferret O, et al. Searching news articles using an event knowledge graph leveraged by Wikidata[C]. In Companion Proceedings of the 2019 World Wide Web Conference (WWW '19). Association for Computing Machinery, New York, NY, USA, 2019: 1232-1239.
- [63] Rony M, Chaudhuri D, Usbeck R, et al. Tree-KGQA: An unsupervised approach for question answering over knowledge graphs[J]. IEEE Access, 2022, 10: 50467-50478.
- [64] Zhang Y, Li P, Liang H, et al. Fact-tree reasoning for n-ary question answering over knowledge graphs[C]. In Findings of the Association for Computational Linguistics, Dublin, Ireland, 2022: 788-802.
- [65] Saxena A, Kochsiek A, Gemulla R. Sequence-to-sequence knowledge graph completion and question answering[C]. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics, Dublin, Ireland, 2022, 1: 2814-2828.
- [66] Jiang Z, Chi C, Zhan Y. Research on medical question answering system based on knowledge graph[J]. IEEE Access, 2021, 9: 21094-21101.
- [67] Pan H, Yang X. Intelligent recommendation method integrating knowledge graph and Bayesian network[J]. Soft Computing, 2023, 27(1): 483-492.
- [68] Muto K, Oyama S, Noda I. Explainable recommendation using knowledge graphs and random walks[C]. IEEE International Conference on Big Data (Big Data), Osaka, Japan, 2022: 4028-4023.
- [69] Zeng Y, Du J, Xue Z, et al. Scientific and technological news recommendation based on knowledge graph with user perception[C]. IEEE 8th International Conference on Cloud Computing and Intelligent Systems, CCIS, Chengdu, China, 2022: 491-495.

- [70] Bordes A, Usunier N, García-Durán A, et al. Translating embeddings for modeling multi-relational data[C]. In Proceedings of the 26th International Conference on Neural Information Processing Systems, 2013: 2787-2795.
- [71] Nickel M, Tresc V, Krieger H, et al. A three-way model for collective learning on multi-relational data[C]. Proceedings of the 28th International Conference on Machine Learning, ICML, Bellevue, Washington, USA, 2011: 809-816.
- [72] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space[C/OL]. 1st International Conference on Learning Representations, ICLR, Scottsdale, Arizona, USA, 2013, <https://arxiv.org/pdf/1301.3781.pdf>.